

Mahrholz, Gaby (2018) *Implicit perceived vocal trustworthiness negatively correlates with amygdala activation*. MSc(R) thesis.

<https://theses.gla.ac.uk/30764/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Implicit perceived vocal trustworthiness negatively correlates with amygdala activation

Gaby Mahrholz

A thesis submitted in fulfilment of the requirements for the Degree of
Master of Science (Res)

School of Psychology
College of Science and Engineering
University of Glasgow

October 2017

Abstract

It has long been established that people make rapid judgements about another's personality and that these judgements have lasting influence on subsequent decisions and interactions. Particularly, in voice research, it has been shown that one word of less than 500 ms is sufficient for forming first impressions of a speaker, and that listeners highly agree on who sounds trustworthy, or dominant. Furthermore, the rapid first impressions formed within a listener, still hold true after a prolonged exposure of approximately 3 seconds. It has also been suggested that numerous personality traits can be summarised in a two-dimensional space of trustworthiness, and dominance. Given our intrinsic need of survival, and self-preservation, first impressions are aiding decisions as to whether to approach or avoid a person.

Neurological evidence however has linked activation in the amygdala (more precisely in the superficial (SF) subdivision) to perceived trustworthiness rather than dominance, implying that the amygdala assists in approach/ avoidance decisions but not in identifying whether a person is physically capable of carrying out threatening behaviour. Despite this clear relationship having been extensively researched with face stimuli, the connection between amygdala activation and perceived vocal trustworthiness is poorly understood. Thus, the current study investigated whether amygdala activation correlated with varying levels of vocal trustworthiness. Furthermore, as there has been an ongoing debate as to whether response patterns were linear or quadratic polynomials in face research, a secondary aim of the current study was to explore response patterns.

To achieve that, the study was divided into three experiments. In Experiment 1, vocal word stimuli ('hello') were pre-validated online for perceived trustworthiness, and 15 voices per voice sex were selected for the fMRI experiments. Experiment 2 focussed on recording amygdala activity (in the whole and the SF part of the amygdala) across two implicit task designs – a 1-back task (Experiment 2a) requiring a high level of attention and cognitive load, and a PureTone detection task in which attention and cognitive load were lower. It was hypothesised that amygdala activation would be negatively correlated to perceived vocal trustworthiness in male and female voices, irrespective of task.

Overall, the hypotheses in these experiments were only partially confirmed as significant correlation values were found for male voices but not female voices. Furthermore, results were task dependent with significant results being observed in the high attention/ cognitive load paradigm (Experiment 2a) but not in the PureTone detection task (Experiment 2b). This suggests that the amygdala is sensitive to modulations in socially relevant vocal characteristics related to approach/avoidance decisions, however, a more varied approach of stimuli selection might be required. Given this study was exploratory in nature, these results should be replicated in a confirmatory analysis on an independent data set with more participants. Furthermore, since this study employed univariate methods, multivariate whole brain analysis would aid in establishing additional neural areas involved in processing vocal trustworthiness.

Table of Contents

Abstract	2
Table of Contents	4
List of Tables	6
List of Figures	8
Abbreviations	11
Introduction	12
Perceived trustworthiness and neural activation in the amygdala	15
The current study	20
Experiment 1: pre-validation of stimuli	22
Methods	22
Ethics	22
Participants	22
Stimuli	23
Procedure	23
Data analysis	23
Results	24
Difference between male and female ratings – group level	24
Difference between male and female ratings – individual voice level	25
Calculation of average trustworthiness and selection of voice stimuli for fMRI experiments	27
Differences in the range of average trustworthiness between male and female voice stimuli	27
Discussion	28
Experiment 2a: fMRI experiment – 1-back task	30
Methods	30

Ethics.....	30
Participants.....	31
Stimuli.....	31
fMRI paradigm, and procedure	31
Imaging parameters.....	32
Data analysis.....	33
Results	35
Behavioural results	35
Imaging results – female voices	37
Imaging results – male voices.....	43
Summary of results	49
Experiment 2b: fMRI experiment – PureTone detection task.....	50
Methods	50
Participants.....	50
fMRI paradigm, and procedure	51
Data analysis.....	51
Results	52
Behavioural results	52
Imaging results – female voices	54
Imaging results – male voices.....	60
Summary of results	66
Comparison between tasks – fMRI data.....	67
General discussion and conclusion	69
References	75
Supplementary materials.....	84

List of Tables

Table 1: Welch's t-test for three female voices with significant differences between ratings of female and male listeners	26
Table 2: Spearman's rank-order correlation coefficient for the individual's post-scan ratings and the online validation ratings (Experiment 1)	36
Table 3: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and online validation trustworthiness ratings	39
Table 4: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and post-scan trustworthiness ratings	42
Table 5: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and online validation trustworthiness ratings	45
Table 6: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and post-scan trustworthiness ratings	48
Table 7: Spearman's rank-order correlation coefficient for the individual's post-scan ratings and the online validation ratings (Experiment 1)	53
Table 8: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and online validation trustworthiness ratings	56
Table 9: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and post-scan trustworthiness ratings	59
Table 10: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and online validation trustworthiness ratings	62
Table 11: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and post-scan trustworthiness ratings	65
Table 12: Spearman's rank order correlation coefficients for the group level amygdala activity and the online validation trustworthiness scores (Experiment 1) in the 1-back and the PureTone task, and William's T2 statistic comparing the two tasks.	67

Table 13: Spearman's rank order correlation coefficients for the group level amygdala activity and the post-scan behavioural responses in the 1-back and the PureTone task, and Fisher's r-z transformation comparing the two tasks.....	68
--	----

List of Figures

Figure 1: Boxplot of average trustworthiness ratings for each female and male listeners separately for each voice sex.....	25
Figure 2: Boxplot of the three individual female voices with significant differences between ratings of female and male listeners	26
Figure 3: Range of average trustworthiness for the 15 female and 15 male voices selected for the fMRI experiment	28
Figure 4: Diagram depicting one block of Sparse Sampling paradigm with 15 voice stimuli and one repetition (adapted from Bestelmeyer et al., 2012)	32
Figure 5: Scatterplot of trustworthiness ratings obtained from the participants in the post-scan behavioural experiment, and in the online validation experiment	35
Figure 6: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the female voices	38
Figure 7: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the female voices	41
Figure 8: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the male voices	44
Figure 9: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the male voices	47
Figure 10: Diagram depicting one block of Sparse Sampling paradigm with 15 voice stimuli and one beep (adapted from Bestelmeyer et al., 2012)	51
Figure 11: Scatterplot of trustworthiness ratings obtained from the participants in the post-scan behavioural experiment, and in the online validation experiment	52
Figure 12: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the female voices	55
Figure 13: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the female voices	58

Figure 14: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the male voices	61
Figure 15: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the male voices	64
Figure 16: Female voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1)	85
Figure 17: Female voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1)	86
Figure 18: Female voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings	87
Figure 19: Female voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings	88
Figure 20: Male voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1)	89
Figure 21: Male voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1)	90
Figure 22: Male voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings	91
Figure 23: Male voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings	92

Figure 24: Female voices in the composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1).....	93
Figure 25: Female voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1)	94
Figure 26: Female voices in the composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings.....	95
Figure 27: Female voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings	96
Figure 28: Male voices in the Composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1)	97
Figure 29: Male voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1)	98
Figure 30: Male voices in the Composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings	99
Figure 31: Male voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings	100

Abbreviations

L	Left Hemisphere
R	Right Hemisphere
C	Composite amygdala
SF	Superficial subdivision of the amygdala
R1	Run 1
RC	Runs 1 and 2 Combined

for example:

LCR1 Left Composite Run 1

RSFRC Right superficial amygdala for Runs 1 and 2 Combined

Introduction

It has long been established that people make rapid judgements about another's personality and that these judgements have lasting influence on subsequent decisions and interactions (Allport & Cantril, 1934; Aronovitch, 1976; Borkowska & Pawlowski, 2011; Herzog, 1933; McAleer, Todorov, & Belin, 2014; Oosterhof & Todorov, 2008; Pear, 1931; Scherer, 1972). For example, research has shown that static cues, such as physical appearance, and perceived attractiveness (Efran, 1974; Zebrowitz, 1996) but also dynamic cues like the quality of a handshake, or the regional accent influence who would get hired for a job, or affect the sentence a criminal would receive (Dixon, Mahoney, & Cocks, 2002; Rakic, Steffens, & Mummendey, 2011; Stewart, Dustin, Barrick, & Darnold, 2008). Such first impressions guide our mate choice (Olivola & Todorov, 2010a), and impact on financial decisions (Gorn, Jiang, & Johar, 2008), general political alignment (Olivola & Todorov, 2010b; Tigue, Borak, O'Connor, Schandl, & Feinberg, 2012), and who we vote for (Klofstad, Anderson, & Nowicki, 2015; Klofstad, Anderson, & Peters, 2012; Tigue et al., 2012; Todorov, Mandisodza, Goren, & Hall, 2005). Particularly in voice research, listeners have been shown to make inferences about the speaker's identity, gender, race, and age (Baugh, 2000; Hughes & Rhodes, 2010; Moyse, Beaufort, & Brédart, 2014; Purnell, Idsardi, & Baugh, 1999), or physical attributes like height and weight (Krauss, Freyberg, & Morsella, 2002; Pisanski et al., 2014; Pisanski et al., 2016). Furthermore, just by listening to their voice, rapid judgements are made about the speaker's perceived intelligence (Schroeder & Epley, 2015), confidence levels (Jiang & Pell, 2015), and personality (Allport & Cantril, 1934; Borkowska & Pawlowski, 2011; McAleer et al., 2014). Whether termed as first impressions, thin-slice personality judgements, or zero acquaintance judgements, it is clear that these judgements affect our daily decisions and actions.

Whether the actual veracity of such rapid judgements can be established has been debated (Funder, 2012; Olivola & Todorov, 2010b; Zebrowitz & Collins, 1997; Zebrowitz & Montepare, 2008). Despite evidence suggesting a considerable amount of within-person variability in different photos of the same face (Jenkins, White, Van Montfort, & Burton, 2011), face and voice research have shown high consistency across raters in regards to perceived personality even after brief exposure to a stimulus (see Cronbach's Alpha values between .7 and .98 reported in: McAleer et al., 2014; Mileva, Tompkinson, Watt, & Burton,

2017; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov, Pakrashi, & Oosterhof, 2009; Vernon, Sutherland, Young, & Hartley, 2014; Willis & Todorov, 2006). This suggests exposure to one word (~ 500 ms), or seeing a face for 100 ms is sufficient to form a first impression about a person's personality. It appears that a trustworthy, dominant, or likeable stimulus is perceived as trustworthy, dominant, or likeable consistently across raters (Aronovitch, 1976; Mahrholz, Belin, & McAleer, under review; McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov et al., 2009; Vernon et al., 2014; Willis & Todorov, 2006; Zuckerman & Driver, 1989). This high inter-rater reliability may hint to a kind of universal template of personality traits we establish potentially early on in life, similar to the prototype found for voice identity (Latinus & Belin, 2012), or in Todorov's 'typicality framework' (Todorov, 2012; Todorov, Mende-Siedlecki, & Dotsch, 2013), allowing us to make judgements about a voice or face stimulus in a fraction of a second.

Recent face, and voice research have not only shown that first impressions are formed instantly, and have high consensus between raters, but also that ratings appear stable across various temporal exposure durations (Bar, Neta, & Linz, 2006; Mahrholz et al., under review; McAleer et al., 2014; Oosterhof & Todorov, 2008; Todorov et al., 2009; Willis & Todorov, 2006). McAleer et al. (2014) had 300 participants rating novel speakers saying 'hello' on ten different personality traits, inter alia trustworthiness, and dominance. In a follow up study, Mahrholz et al. (under review) found moderate to strong correlations between trait judgements after one word (< 500ms), and a sentence (~3 sec). Taken together these findings suggest that an exposure to the vocal stimuli of an average duration of around 500 ms is sufficient to make reliable judgements about the speaker, and that these first impressions formed within a listener still hold true after a prolonged exposure of approximately 3 seconds. Similar results were reported for several personality traits in research employing face stimuli; Willis and Todorov (2006) found medium to high correlation values between ratings made after 100, 500, and 1000 ms of exposure, and the ones obtained without time constraints. Similarly, Bar et al. (2006) reported moderate correlation coefficients between 39 and 1700 ms of stimuli exposure. Despite recruiting different sets of raters for each exposure condition, these results suggest that, similar to voice research, first impressions of trait ratings from faces made after short exposure are

still valid after prolonged exposure to the same stimulus, implying stability across temporal duration.

Explanations, as to whether forming first impressions quickly is beneficial, can be found in Evolutionary Theory, which proposes that traits deemed for immediate survival or self-preservation might be judged more instantly than others (trustworthiness, dominance, threat vs intelligence; Bar et al., 2006; Todorov et al., 2009). The evaluation whether a person is threatening to our health, influences the decision to approach, or avoid that person. Altering the approach/ avoidance angle, Oosterhof and Todorov (2008) proposed we use overgeneralisations, that is to say momentary, dynamic emotion perceptions, to make inferences about static personality traits such as trustworthiness, or dominance. The notion of overgeneralisation is not particularly new, however Oosterhof and Todorov (2008) extend on and define the previous concept further. In their opinion, valence evaluation links to threat as an assessment as to whether to approach or avoid a person, and the evaluation of dominance relates to the physical strength of the person and therefore their ability to carry out the threat. Both models hinge around the approach/ avoidance theorem, but whereas evolutionary explanations focus on survival, overgeneralisation sees the basis for judging personality traits in affect.

Furthermore, potentially in order to assist us making rapid approach or avoidance decisions about a person, numerous personality traits can be summarised in a two-dimensional space via principal component analysis; the first component is commonly related to valence (Oosterhof & Todorov, 2008; Sutherland et al., 2013), frequently aligned to traits of trustworthiness (McAleer et al., 2014), integrity (which includes trustworthiness; Tigue et al., 2012), or likeability (Zuckerman & Driver, 1989), whilst component two concerns dominance (McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zuckerman & Driver, 1989), or physical prowess (which includes dominance; Tigue et al., 2012). Oosterhof and Todorov (2008) have shown threat judgements to correlate negatively with valence and PC1, and positively with dominance and PC2 from face stimuli. However recent neuroimaging research (Todorov & Engell, 2008) suggests that amygdala activation is correlated to trustworthiness/ valence (PC1) rather than dominance (PC2). According to Oosterhof and Todorov's (2008) overgeneralisation model, this would imply that the amygdala assists in making rapid decisions as to whether to approach or avoid a person, but

not to assess whether a person possesses the physical strength to carry out threatening behaviour.

Perceived trustworthiness and neural activation in the amygdala

Neurological evidence suggests the amygdala plays a key role in the perception of trustworthiness. The amygdala was a region thought of originally as the early threat detection centre of the brain (LeDoux, 2003, 2012) but has since been shown relevant in numerous affective processes, such as consolidation of emotional memory (McGaugh, 2004; Roozendaal, McEwen, & Chattarji, 2009), emotional attention (Vuilleumier, 2005), and relevance detection (Sander, Grafman, & Zalla, 2003; Zalla & Sperduti, 2013). Early evidence that the amygdala is involved in perceived facial trustworthiness originate from human lesion research. Adolphs, Tranel, and Damasio (1998) found that patients with bilateral amygdala damage (including patient SM) perceived untrustworthy-looking faces as more trustworthy compared to healthy, and brain-damaged controls. Surprisingly, participants with uni-lateral amygdala damage did not differ from controls. SM also exhibited a preference for closer interpersonal space which suggests a possible impairment of the approach/ avoidance mechanisms (Kennedy, Gläscher, Tyszka, & Adolphs, 2009). Kosciak and Tranel (2011) explored the role of the amygdala in relation to interpersonal trust in a multi-round, multiplayer economic trust game, and found that patients with unilateral amygdala damage showed increased benevolent behaviour, more specifically they were inclined to increase trust after experiencing betrayal. Brain-damaged controls responded indifferent to either trustworthy behaviour, or betrayal, whereas neurologically healthy controls employed a tit-for-tat strategy, meaning trust was repaid by trust, betrayal by betrayal. Conversely, potentially due to an intact amygdala, some patients with prosopagnosia were able to perceive trustworthiness without having the ability to recognise faces (Quadflieg, Todorov, Laguesse, & Rossion, 2012; Todorov & Duchaine, 2008). This early research displays clearly that perceptions of trustworthiness are impaired once the amygdala is damaged either unilaterally or bilaterally.

Thereupon, numerous neuroimaging studies have been conducted strengthening the early connection between amygdala activation and perceived trustworthiness. Winston et al.

(2002), utilising an implicit (decision age: high school vs university student) as well as an explicit (trustworthy vs untrustworthy face) task, reported an increase in bilateral amygdala activation for untrustworthy faces independent of task. Given that implicit and explicit runs were counterbalanced in Winston et al. (2002), Engell, Haxby, and Todorov (2007) proposed to employ implicit tasks solely, due to a potential priming effect in participants receiving the explicit prior to the implicit tasks. They remarked that the explicit trustworthiness instructions could still influence the implicit age assessment ratings which, in turn, could lead to an overestimation of amygdala involvement during the implicit runs. However, Winston et al.'s (2002) results of a negative correlation between perceived facial trustworthiness and amygdala activation were replicated for "pure" implicit trials (Engell et al., 2007; Todorov & Engell, 2008; Todorov, Said, Oosterhof, & Engell, 2011), explicit trials (Said, Dotsch, & Todorov, 2010; Todorov et al., 2011), and with computer-generated faces (Todorov, Baron, & Oosterhof, 2008). More recently, Freeman and colleagues (2014) investigated the neural basis of trustworthiness outside conscious awareness based on previous behavioural findings regarding minimal exposure (Bar et al., 2006; Todorov et al., 2009; Willis & Todorov, 2006). Participants viewed stimuli for 33ms whilst completing an implicit 1-back task. They reported a negative relationship between amygdala activation and trustworthiness, and claimed the amygdala coded trustworthiness even before face stimuli were consciously perceived. Results held true across event-related and block designs, using both real and computer-generated face stimuli.

Two meta-analysis tried to assess the neural networks involved in social evaluation of faces. Both included studies investigating either trustworthiness or attractiveness, firstly due to the high correlations with one another, and their respective representation in PC1 (0.60 to 0.80; Oosterhof & Todorov, 2008; Todorov, Baron, et al., 2008; Todorov, Said, Engell, & Oosterhof, 2008), and secondly for conceptual reasons: both trustworthiness and attractiveness convey fundamental information for succeeding the social world; the former with regards to modulating approach and avoidance behaviour towards strangers, the latter in connection with mate selection. Employing different inclusion criteria, and different analysis techniques (Activation-Likelihood Estimation (ALE; Bzdok et al., 2011), Multi-level Kernel Density Analysis (MKDA; Mende-Siedlecki, Said, & Todorov, 2013)), both identified activations pattern in the amygdala in relation to facial trustworthiness – Mende-Siedlecki in

the right amygdala, and Bzdok et al. (2011) in the bilateral amygdala, more specifically the superficial (SF) part. However, Mende-Siedlecki et al. (2013) did not distinguish between explicit and implicit studies. Adding this additional comparison, Bzdok et al. (2011) reported bilateral amygdala activation of the laterobasal (LB), and SF subdivisions for implicit tasks, and bilateral SF amygdala activation for the explicit tasks. However, this was done for a combined analysis of trustworthiness and attraction due to the small number of studies involved. Whilst it seems sensible to merge attractiveness and trustworthiness studies to study complex social behaviours, it makes untangling results for specific traits like trustworthiness rather difficult. Whether a difference in neural activation pattern exists between explicit and implicit task designs, it is obvious that the amygdala is involved in the perception of trustworthiness from faces.

A further aspect to discuss is the ambiguity in response pattern which has received a fair representation in the literature so far. Initial studies reported negative linear relationships (Engell et al., 2007; Mende-Siedlecki et al., 2013; Todorov & Engell, 2008; Winston et al., 2002), indicating that amygdala activation increases with decreasing trustworthiness, whereas others (Said, Baron, & Todorov, 2009; Said et al., 2010; Todorov et al., 2011) have described non-linear/ quadratic relationships, with extreme trustworthiness values (either positive or negative) eliciting higher amygdala activation than centre values of the continuum. Freeman et al. (2014) and Todorov, Baron and Oosterhof (2008) found both linear and quadratic effects to trustworthiness in the amygdala depending on either the type of fMRI design (linear in block design, quadratic in event-related design; Freeman et al., 2014), or in respect to hemisphere (linear in right amygdala, quadratic in left amygdala; Todorov, Baron, et al., 2008). Mattavelli, Andrews, Asghar, Towler, and Young (2012) aimed to address the contradiction of the activation pattern using morphing techniques of face prototype stimuli (previously created via photograph averaging technique) along the independent dimensions of gender and trustworthiness. They employed a block-design fMRI paradigm, and instructed participants to look at the images and press a button when a small red spot appeared (red spot detection task). The study discovered quadratic responses in the bilateral amygdala, and other core face-selective regions in relation to trustworthiness, and gender. Said, Dotsch, and Todorov (2010) sought to explain the discrepancy in the literature with the concept of face typicality. The authors observed a linear pattern when

real face stimuli were used whereas quadratic responses emerged when studies used artificially created faces. After re-validating the initial stimuli from previous studies on face typicality, they found a linear responses pattern between amygdala activation and valence, when valence and typicality were correlated linearly. In studies reporting quadratic responses, the face stimuli showed a quadratic pattern between valence and typicality as well. This explanation neither invalidates the linear nor the non-linear activation pattern, and indicates that the amygdala is activated more strongly the further stimuli are removed from a prototype, similar to the one proposed by Latinus and Belin (2011) for voice identity.

Despite extensive research showing a relationship between amygdala activation and perceived facial trustworthiness (or valence), there is little research on the neural networks of perceived vocal trustworthiness. One study to date has investigated neural correlates of social judgements using voice stimuli (Hensel, Bzdok, Mueller, Zilles, & Eickhoff, 2015). The authors recorded brain activation whilst participants judged vocal stimuli on the social traits of attractiveness and trustworthiness, as well as emotion (happiness), and identity trait (age). Stimuli were sentences frequently occurring in everyday social interaction (average duration ca 2.5 sec) spoken by 40 German native speakers between 20 and 83 years of age. In total there were 5 blocks with 8 trials each for each test category, and the question (how trustworthy, how attractive, how happy, how old) did not alter within blocks. The rating scale, from 1 [not at all] to 8 [very], required buttons responses covered by the 4 long fingers on each hand. Neural activity in relation to the social trait judgements, i.e. attractiveness and trustworthiness, were found in the left superior frontal gyrus (SFG), bilateral Inferior Parietal Cortex (IPC), and dorso-medial PFC (dmPFC) extending into perigenual anterior cingulate cortex (pACC). The authors did not report individual brain regions for trustworthiness, and attractiveness separately.

It can be seen that Hensel et al. (2015) report no evidence of the amygdala's involvement in vocal social trait judgements (no mention). This could be due to various reasons. Firstly, Hensel et al. (2015) did not control for listeners' age; their participants' age ranged from 21 to 60 years with an average age of 36. Behaviourally, Castle et al. (2012) indicated a shift in trustworthiness perceptions with age, i.e. older adults become more trusting, which could influence results. Furthermore, given a vast literature analysing perceptual, cognitive, and neural changes with age (see for example Grady, 2012; Gutchess, 2014; Nyberg, Lovden,

Riklund, Lindenberger, & Backman, 2012; Reuter-Lorenz & Park, 2010; Salthouse, 2010; St Jacques, Dolcos, & Cabeza, 2009) but also compensation strategies of the brain, like the 'compensation related utilization of neural circuits hypothesis' (CRUNCH; Reuter-Lorenz & Cappell, 2008), hemispheric asymmetry reduction in older adults (HAROLD; Cabeza, Anderson, Kester, & McIntosh, 2002), or the posterior-anterior shift in ageing (PASA; Davis, Dennis, Daselaar, Fleck, & Cabeza, 2008), there is a possibility to miss brain regions involved in younger adults only by not controlling for age. A second reason could be the selection of sentence stimuli. As the amygdala could be responsible for automatic rapid first impression of relevant stimuli (LeDoux, 2007; Sander et al., 2003), for example how threatening or trustworthy a person is, sentences might be too long of a duration to elicit activation in the amygdala. Hensel et al. (2015) reported significantly longer reaction times for trustworthiness than age, happiness, or attractiveness, suggesting the identified brain regions could be involved in higher cognitive processes such as decision-making rather than initial assessments of trustworthiness. A further reason as to why the amygdala found no mentioning in Hensel et al. (2015) could be the authors' combined analysis of male and female voice stimuli. Given the close connection between the concepts of trustworthiness and threat (Oosterhof & Todorov, 2008), and that males are perceived as more threatening than females in an evolutionary sense (Puts, 2010, 2016; Puts, Apicella, & Cárdenas, 2012), it is possible that amygdala activation relates to male but not female voices. This would be missed if voice sexes were not analysed separately. Another reason could be found in the task design. Hensel et al. (2015) used an explicit task. As it was not entirely clear whether there was a distinction in neural activation pattern between explicit and implicit tasks in face research, the task choice might contribute to determined activation patterns in the brain, or a lack thereof. Lastly, there might be no activation in the amygdala from vocal stimuli in relation to social judgements: We are living in a predominantly visual world, and audio-visual research shows that trustworthiness appears to be driven by an integration of either both modalities, or mainly the face dimension (Mileva et al., 2017; Rezlescu et al., 2015) which could explain tracing perceived facial trustworthiness in the amygdala but not vocal trustworthiness.

The current study

This paper thus investigates whether amygdala activation correlates with varying levels of vocal trustworthiness and whether it is influenced by experimental task. Focus will be on the personality trait trustworthiness due to being one of the key traits emerging from Principal Component Analysis (PCA). Trustworthiness appears to be a more stable trait than dominance as perception and preferences for dominance might shift during the ovarian cycle (DeBruine et al., 2010; Jones et al., 2008). Additionally, in face research perceived trustworthiness rather than dominance has been linked to activation in the amygdala, particularly to superficial areas (SF; Bzdok et al., 2011; Mende-Siedlecki et al., 2013; Todorov & Engell, 2008). As findings from voice research frequently mirror findings from face research (Yovel & Belin, 2013), the bilateral “whole” amygdala (henceforth referred to as ‘composite amygdala’), as well as its superficial (SF) subdivision were selected as regions of interest (ROIs) for this study.

The only voice study so far exploring the neural correlates of social judgments did not report the amygdala as one of the key regions in regards to trustworthiness, however this might potentially be due to employing an explicit task, using sentence stimuli, participants’ age, and/ or combining analysis for male and female voices. Therefore, this study employs an implicit task design and the socially relevant word stimulus ‘hello’. As the stability of trait ratings along varying temporal durations has been established (Mahrholz et al., under review), selecting a word allows to investigate first impressions of perceived personality rather than potentially tapping into processes of decision-making. A further aspect to consider is age of the participants given research indicating a shift of trustworthiness perception, and/or a general change of neural networks in older participants (Castle et al., 2012). To avoid age being a confounding factor, the age range for this experiment will be kept consistently between 17 and 30 years for all project stages. Moreover, this study will analyse female and male voice voices separately.

To answer the overall research question of the correlation between amygdala activation and perception of trustworthiness in voices, the study was divided into three experiments: Experiment 1 pre-validated voice stimuli that had previously been used by our lab (Mahrholz et al., under review; McAleer et al., 2014) online in order to select the 15 male and 15 female voice stimuli with varying levels of vocal trustworthiness for the fMRI experiment.

Experiment 2 focussed on recording amygdala activation in regards to implicit perceived trustworthiness from voices across two fMRI experiments. The 1-back task (Experiment 2a), a design that has previously been used in face research (Freeman et al., 2014), requires participants to directly attend to the voice stimuli. As it requires memory of the previous voice to compare between the current and previous stimulus, cognitive load might interfere with the social perception this study intends to investigate. The PureTone detection task (Experiment 2b) was designed paralleling the red spot detection task used by Mattavelli et al. (2012). It is potentially easier for the participants as it is less demanding on memory and attention as participants are only required to determine whether the current stimulus is a voice or not. Subsequent to scanning, participants were completing a post-scan behavioural experiment in which their perceived trustworthiness of each voice stimulus will be gathered. As Engell et al. (2007) established that behavioural consensus data of trustworthiness were a better predictor of amygdala activation than individual ratings, amygdala activation will be correlated to trustworthiness ratings obtained in online validation experiment (Experiment 1) and, as a form of validation, to those from the post-scan tasks. It is expected that amygdala activation will be negatively correlated with varying levels of vocal trustworthiness in female and male voices irrespective of task. Similarly to Mattavelli et al. (2012), potential responses will be investigated for the type of their correlation, whether a quadratic, or linear model would describe the relationship best. Findings from this study will allow to draw conclusions to previous voice, and face research as to whether amygdala activation is related to perceived trustworthiness. If findings from face research hold true for voices, this would strongly imply modality-independent underpinnings of trustworthiness judgements.

Experiment 1: pre-validation of stimuli

The purpose of this online validation experiment is to select the voice stimuli which to include in the fMRI experiments (2a, and 2b). As this study aims to establish brain regions involved in trustworthiness perceptions from vocal stimuli, it is important to determine trustworthiness ratings for each of the voice stimuli prior to selecting a subset for the fMRI experiment. Furthermore, as McAleer et al. (2014) and Mahrholz et al. (under review) used different scales (Likert vs VAS scale) in their respective experiments, revalidation of the stimuli became necessary. However, Experiment 1 also seeks to investigate inter-rater reliability, examine whether there are differences between the behavioural responses of male and female listeners on a group, as well as on an individual voice level, describe how the overall trustworthiness for each is calculated, and establish whether male and female voices differ in the range of perceived trustworthiness.

Methods

Ethics

Ethics was granted for the pre-validation online experiment by the University of Glasgow Ethics Committee whose guidelines are in accordance with the ethical standards in the Declaration of Helsinki (1964). Participant gave consent when signing up on the University of Glasgow Psychology online Experiment website (<http://experiments.psy.gla.ac.uk/index.php>) and clicking the button to start the experiment. They agreed to have the right to withdraw from the experiment at any times and without particular reason, were guaranteed anonymity, confidentiality, and secure storage of their data.

Participants

50 participants (35 female: 19.9 ± 2.52 years (range: 18-27 years); 15 males: 18.9 ± 1.61 years (range: 18-24 years)) took part in the online rating experiment. Participant recruitment was via the University of Glasgow School of Psychology Subject Pool. Advertising was placed for participants between 17 and 30 years of age. All first year Psychology students at the University of Glasgow were reimbursed one participation credit

as part of their undergraduate degree. No monetary incentive was given for taking part in the experiment.

Stimuli

Stimuli recordings of the word 'hello' (16 bit mono, 44100Hz, WAV format), that had been used in two previous studies from our lab (Mahrholz et al., under review; McAleer et al., 2014), were selected for the online rating experiment. For detailed recording and extracting procedures refer to McAleer et al. (2014) or Mahrholz et al. (under review). All 45 female (21.6 ± 3.80 years; age range: 17-30) and 43 male speakers (average age of 23.6 ± 3.36 years; age range: 17-30) were Scottish, and between 17 and 30 years of age at the time of recording. The stimuli had an average duration of 378 ± 60.31 ms, and 388 ± 65.12 ms for female and male voices respectively. As different intensity levels can influence perceptions of personality traits (Dunbar & Burgoon, 2005), stimuli were normalised for intensity through Matlab (The MathWorks, Inc., Natwick, Massachusetts, USA).

Procedure

Participants gave consent by signing up on the School of Psychology Experiment website (<http://experiments.psy.gla.ac.uk/index.php>), and starting the experiment. They were instructed to complete the experiment in a quiet environment using headphones or speakers. Each participant was asked to listen to the stimuli, and subsequently provide a rating on a scale from 1 [extremely untrustworthy] to 7 [extremely trustworthy] of how trustworthy they perceived that voice to be. The 88 voice stimuli (45 female, 43 male) were blocked by gender, and each participant rated both blocks in a counterbalanced order either starting with female, or male voices. Each voice was heard twice within each block, resulting in a total of 176 ratings per participant. There were no breaks between voices within a block, however participants were provided with an untimed break between the two blocks. The entire experiment lasted for approximately 15 minutes per participant.

Data analysis

Additionally to the age criterion (participants had to be between 17 and 30 years of age), the following exclusion criteria were specified prior to commencing the experiment to try removing participants not taking the online task seriously: 1) 75% of all given responses could not be on a single rating (i.e. rating each voice as a 4), and 2) two thirds of the second

ratings had to fall within 2 rating points of the first (i.e. if first rating is a 3, the second had to be between 1 and 5). These criteria resulted in the exclusion of 6 subjects of the initially tested 56, leaving the ratings of 50 participants (35 females, 15 males) included in the study for further analysis.

Cronbach's alpha was employed as a measure of inter-rater reliability to establish a level of consistency of perceptions of trustworthiness between the 50 listeners. The Cronbach's alpha scores were high with .94 for female voices, and .95 for male voices.

The average of listener's first and second rating was taken as the final trustworthiness rating. Trustworthiness scores for each voice were then calculated as the average of all female, and male listeners, and a Welch's two-sample t-test was subsequently employed to investigate whether there was a significant difference between the ratings of male and female participants. Furthermore, individual voices were checked for significant differences in the ratings between male and female listeners. Next, a Levene's test was computed to explore whether the range of trustworthiness ratings between female and male voices was significantly different. All critical significance levels were set to .05.

Results

Difference between male and female ratings – group level

A Welch's two-sample t-test revealed no significant differences between male and female listeners for either female, or male voices (see Figure 1). Female voices: $M_{\text{FemaleListeners}} = 4.17$, $SD_{\text{FemaleListeners}} = .51$; $M_{\text{MaleListeners}} = 4.16$, $SD_{\text{MaleListeners}} = .57$; $t(86.692) = .077$, $p = .939$; Male voices: $M_{\text{FemaleListeners}} = 4.12$, $SD_{\text{FemaleListeners}} = .57$; $M_{\text{MaleListeners}} = 4.16$, $SD_{\text{MaleListeners}} = 0.59$; $t(83.937) = -.305$, $p = .761$.

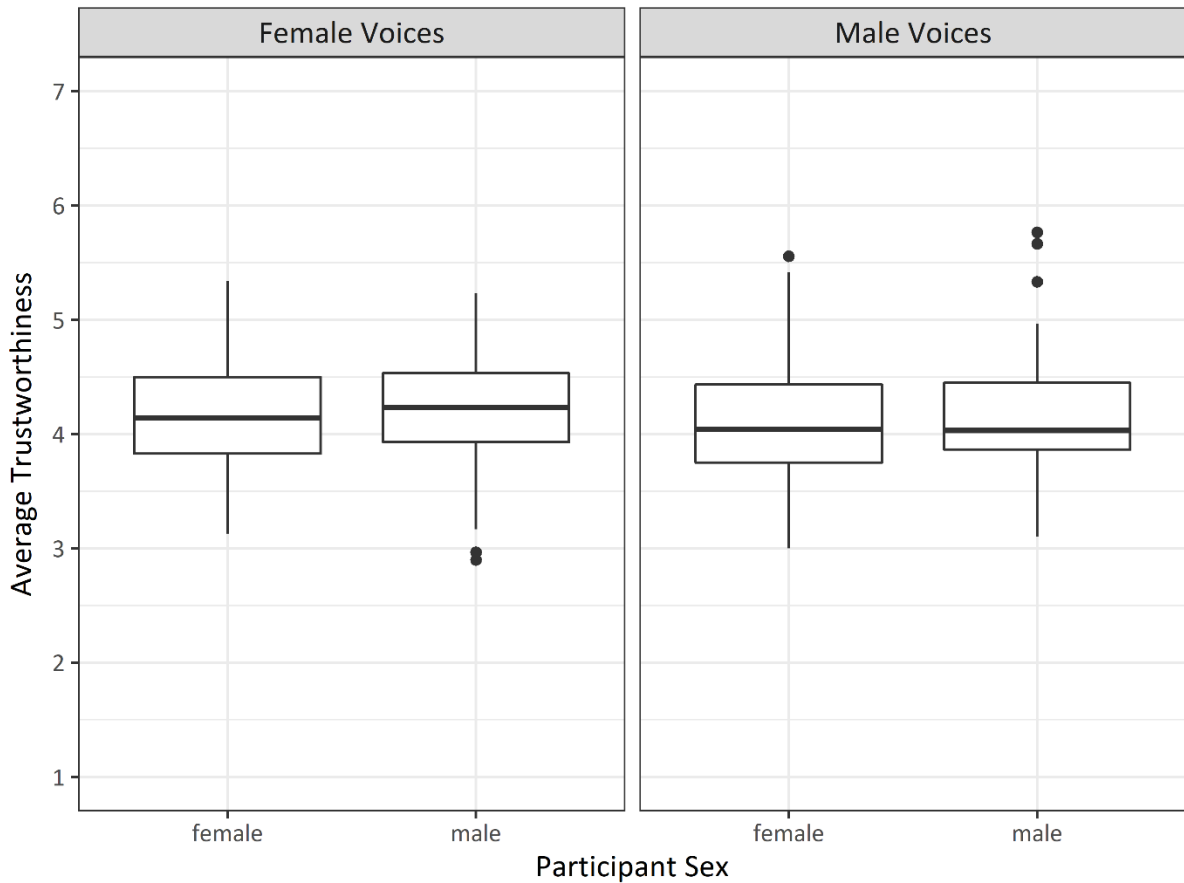


Figure 1: Boxplot of average trustworthiness ratings for each female and male listeners separately for each voice sex

On further inspection of the data, seven outliers were identified (see Figure 1). Three female voices were rated low by male participants that they classified as outliers. Within the male voices, four outliers were identified on the high rating end of the scale – one on the female and three on the male participants' rating scale. However, non-significant differences between male and female listeners' perception of male and female voices persisted after the seven voice outliers were removed from the data set (Female voices: $M_{\text{FemaleListeners}} = 4.17$, $SD_{\text{FemaleListeners}} = .51$; $M_{\text{MaleListeners}} = 4.25$ $SD_{\text{MaleListeners}} = .49$; $t(84.929) = -.734$ $p = .465$; Male voices: $M_{\text{FemaleListeners}} = 4.03$, $SD_{\text{FemaleListeners}} = .53$; $M_{\text{MaleListeners}} = 4.16$, $SD_{\text{MaleListeners}} = 0.44$; $t(83.937) = -.305$, $p = .761$). Therefore no outliers were excluded from the data set.

Difference between male and female ratings – individual voice level

A Welch's two-sample t-test was performed for each individual voice to analyse whether male and female listeners perceived them differently. Male and female listeners differed in their perception of three female voices (see Table 1, and Figure 2) which were henceforth

excluded from the data set. This resulted in a total of 42 female, and 43 male voices for stimuli selection.

Table 1: Welch's t-test for three female voices with significant differences between ratings of female and male listeners

Voice	Average trust female listeners	Standard deviation female listeners	Average trust male listeners	Standard deviation male listeners	t	df	p value
RWF24	3.66	1.07	2.97	0.90	2.352	31.490	0.025
F16	4.81	0.88	4.33	0.59	2.270	38.827	0.029
F30	4.11	0.98	4.70	0.88	-2.081	29.291	0.046

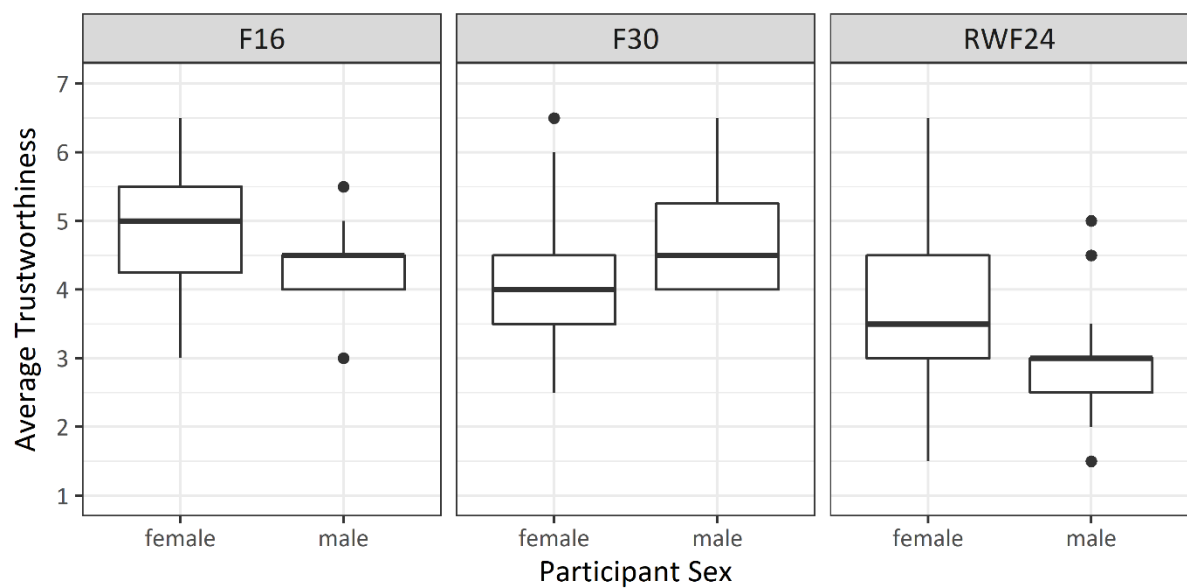


Figure 2: Boxplot of the three individual female voices with significant differences between ratings of female and male listeners

Calculation of average trustworthiness and selection of voice stimuli for fMRI experiments

As the previous analysis revealed no significant differences between the perception of vocal trustworthiness between the male and female listeners, the final trustworthiness rating for each voice was calculated as the average of male listeners rating and female listener.

The 42 female, and 43 male voice stimuli were each arranged by ascending trustworthiness scores. For the fMRI experiment, 15 voices were selected per voice sex in approximately equal trustworthiness distance from one another. The average distance of trustworthiness between voices was approximately 0.16 for female voices, and 0.18 for male voices.

Differences in the range of average trustworthiness between male and female voice stimuli

Overall, perceived trustworthiness in female voices ranged from 3.05 to 5.27, and in the male voices from 3.05 to 5.61 (see Figure 3). In order to assess whether the range of the average trustworthiness scores differed between male and female voices, Levene's test was conducted once for the full data set, and again for the subset of stimuli selected for the fMRI experiments. In both cases Levene's test showed no significant differences of perceived trustworthiness range between male and female voices (full data set: $F(1,83) = .080$, $p = .778$; fMRI-subset: $F(1,28) = .403$, $p = .531$).

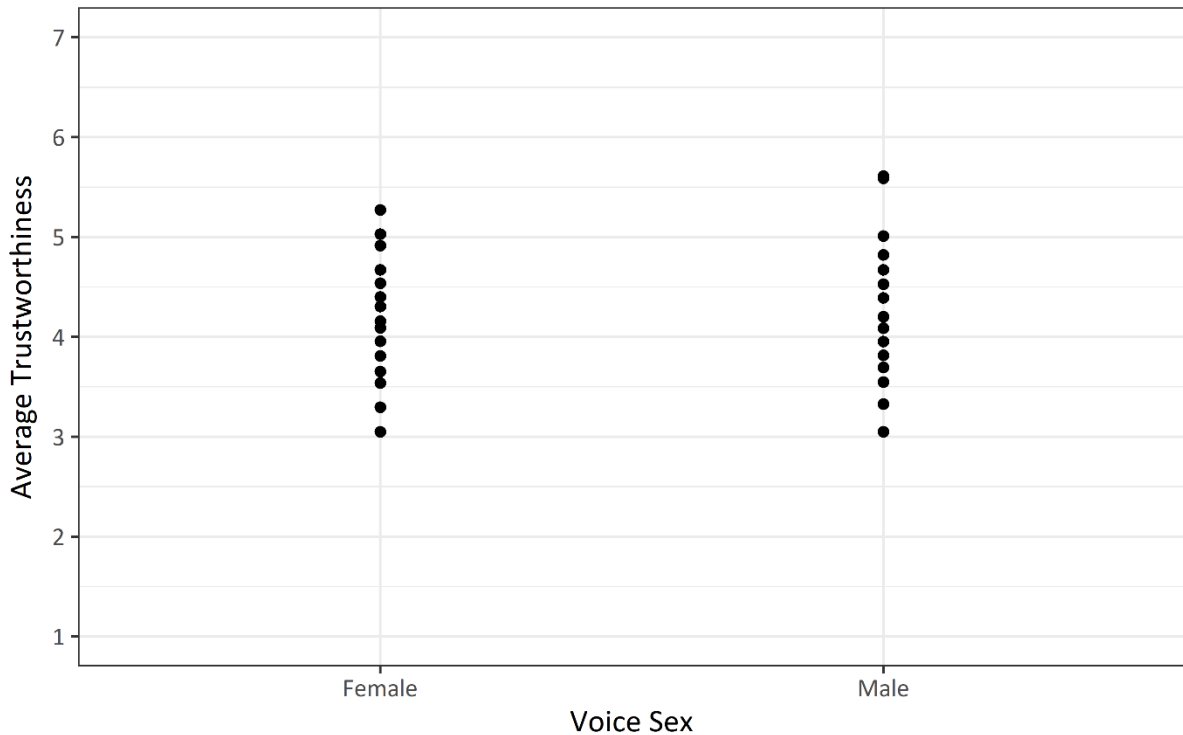


Figure 3: Range of average trustworthiness for the 15 female and 15 male voices selected for the fMRI experiment

Discussion

The aim of Experiment 1 was to select 15 male, and 15 female voice stimuli to be used in the follow-up fMRI experiments (2a, and 2b). Previous to the selection process, inter-rater reliability was computed, determining whether differences of perception in trustworthiness existed between female and male listeners, and established whether the range of trustworthiness ratings differed between female and male voices in the full data set as well as the final selection of voices for the fMRI experiment.

As a measure for inter-rater reliability, high Cronbach's alpha values were obtained showing that listeners agreed strongly on which female and male voices sound trustworthy. This compares favourably to other studies that also achieved high agreement amongst their raters in face and voice research (compare to Mahrholz et al., under review; McAleer et al., 2014; Oosterhof & Todorov, 2008; Willis & Todorov, 2006). High agreement between raters for perceived trustworthiness may hint to an internalised universal representation, and thus to prototypical coding, similar to the one for voice identity reported by Latinus and Belin (2011).

Furthermore, no significant differences in perceived trustworthiness between male and female listeners were found for male and female voices. This is in accordance with previous research (Bruckert et al., 2010; Mahrholz et al., under review) showing that a trustworthy voice is perceived as trustworthy regardless of listeners' sex. Therefore, the overall average trustworthiness for each voice was calculated with female and male judgements contributing in equal parts. This was done to ensure that male and female listeners contributed to the same extent to the overall trustworthiness score of each voice despite having fewer male listeners completing the experiment. Though, on an individual voice level, male and female listeners differed in their perception of trustworthiness for three female voices which were excluded from the sample prior to selecting the final stimuli set of 15 male and 15 female voices for the fMRI experiment. The subset of 15 voice stimuli per voice sex were selected from the full data set in approximately equal trustworthiness distance from one another.

Additionally, it was established that the trustworthiness range of female and male voices was not significantly different on both the full data set and the subset. This shows firstly a good range of voices varying in degree of trustworthiness, and secondly that neither male, nor female voices had extreme trustworthiness rating or were pooled at the high or low extreme of the scale.

Experiment 2a: fMRI experiment – 1-back task

The aim of Experiment 2a was to determine whether amygdala activation correlates with varying levels of vocal trustworthiness. As facial trustworthiness has been linked to amygdala activity, particularly in the SF subdivision (SF; Bzdok et al., 2011; Mende-Siedlecki et al., 2013; Todorov & Engell, 2008), both the bilateral SF and the bilateral composite amygdala were selected as regions of interest for this study. Recordings of the word ‘hello’ from 30 different speakers (15 female), pre-validated in Experiment 1, were used as vocal stimuli to investigate initial first impressions of perceived trustworthiness. Contrary to Hensel et al. (2015) who used an explicit task, the current study employed an implicit design. A 1-back task was created to maintain participants’ attention to the voices. Cognitive load is high in this task as it requires memory of the previous voice to make a comparison between the current and previous stimulus. Behavioural perceptions of trustworthiness were gathered subsequent to the fMRI experiment. As behavioural research report a shift in perceptions with age (Castle et al., 2012), the current research study keeps age consistent between 17 and 30 years. Furthermore, female and male voices were analysed separately for evolutionary reasons (Puts, 2010, 2016; Puts et al., 2012), given the close connection between trustworthiness and threat (Oosterhof & Todorov, 2008). Amygdala activation was then correlated to the consensus data of Experiment 1, and the post-scan behavioural results, due to consensus data of trustworthiness being a better predictor of amygdala activation than individual ratings (Engell et al., 2007). A further aim of this study was to investigate whether the pattern of response was linear or quadratic (see Bzdok et al., 2011; Freeman et al., 2014; Mattavelli et al., 2012).

Methods

Ethics

Ethics was granted for the fMRI and subsequent behavioural task by the University of Glasgow Ethics Committee whose guidelines are in accordance with the ethical standards in the Declaration of Helsinki (1964). For the fMRI part of the experiment, participants gave signed consent to acknowledge they have the right to withdraw from the experiment at any times, were guaranteed anonymity, confidentiality, and secure storage of their data.

Additionally, they underwent thorough eligibility checks prior to being scanned. As the fMRI tasks were orthogonal in nature, additional written consent was obtained before the start of the behavioural experiment, giving people the opportunity to opt out of the subsequent task.

Participants

Participants were recruited via the University of Glasgow School of Psychology Subject Pool. Advertising was placed for native speakers between the 17 and 30 years of age, without a diagnosis of mental disorders, not on any medication, without hearing impairments, and not having participated in the validation stage of the experiment (Experiment 1). Twenty participants (10 male, average age: 21.25 ± 3.43 , age range: 17-29) took part in the fMRI experiment. An additional five participated but were excluded during data analysis (see Exclusion criteria for fMRI participants, page 33). Participants received £15 for participating in the experiment.

Stimuli

The 15 female, and 15 male voice stimuli that had been pre-selected in Experiment 1 were used in the fMRI experiment. The female voices had an average age of 20.9 ± 4.08 years (age range: 17-30), and an average duration of 387.42 ± 56.81 ms, whereas the male voices had an average age of 22.9 ± 3.69 years (age range: 17-30), and an average duration of 372.72 ± 53.08 ms.

fMRI paradigm, and procedure

In order to allow for the clear presentation of auditory stimuli without interference by scanner noise (Perrachione & Ghosh, 2013), a Sparse Sampling (or ‘Sampling with holes’) paradigm was used for functional runs in this experiment as it introduces delay after each functional volume acquisition. Stimuli were presented via headphones at approximately 85db. Participants completed 4 functional runs (2 male, 2 female) with 10 blocks each. Each block consisted of 15 voice stimuli and either one, or two repetitions. Blocks were separated by 15 seconds of silence to allow for the hemodynamic response function to level back to baseline. The order of the runs (male-male-female-female; or female-female-male-male), the position of the repetitions, and the order of the voice stimuli within each block was

counterbalanced. Figure 4 displays one block of the Sparse Sampling paradigm. Stimuli onset occurred within a 1 second gap, with no silent periods during a block.



Figure 4: Diagram depicting one block of Sparse Sampling paradigm with 15 voice stimuli and one repetition (adapted from Bestelmeyer et al., 2012)

Participants completed a 1-back task in which they pressed an allocated button with their right index finger when hearing the exact voice repeated within each block. The lights were turned off and participants were asked to keep their eyes closed. No additional task instructions were given. Subsequent to the 1-back task, an anatomical scan and a voice localiser scan (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Pernet et al., 2015) were obtained, providing a detailed anatomical map of the participant's brain, and highlighting the voice-selective areas in the brain respectively.

After completion of the fMRI procedures, all participants completed a 10-minute behavioural experiment in which they rated each voice stimulus on a Likert scale ranging from 1 [extremely untrustworthy] to 7 [extremely trustworthy]. Equivalent to Experiment 1, the 30 voice stimuli (15 female) were blocked by gender, and each participant rated both blocks in a counterbalanced order. Each stimulus was heard twice within each block. No breaks were included within a block, however participants were able to take a break between the two blocks.

Imaging parameters

Images of blood-oxygenated level dependant (BOLD) signal were acquired using a 3.0T Siemens Tim Trio scanner with a 32-channel head coil. Participants completed four experimental runs (MMFF, or FFMM), each lasting 11.5 min. During the experimental scans, T2*-weighted whole-brain scans were acquired using echo-planar imaging (EPI) [32 interleaved slices, 2mm thickness, 0.2mm gap, 2mm³ in-plane resolution, flip angle 90°, FOV=192, 1151x1152 matrix, TR=3000ms, TA=2000ms, TE=30ms, 230 volumes]. The position of slices was optimised for each participant to ensure recording activity from the temporal lobe (in which TVAs, and amygdala are located).

A whole brain T1-weighted anatomical scan was obtained [192 axial interleaved slices, 1mm thickness, 0.1mm gap, 1mm³ in-plane resolution, flip angle 9°, FOV=25mm, 256x256 matrix, TR=2300ms, TE=2.96ms] using a 3-D magnetization-prepared rapid gradient echo (MPRAGE). The anatomical scan lasted 9 minutes and 50 seconds.

Finally, a 10 min 28 sec voice localizer scan (Belin et al., 2000) was obtained for each participant in order to locate the TVAs. This was a T2*-weighted whole-brain functional scan using EPI [32 interleaved slices, 3mm thickness, 0.3mm gap, 3mm³ in-plane resolution, flip angle 77°, FOV=210mm, 70x70 matrix, TR=2000s, TE=30ms, 310 volumes]. However, the voice localizer is outwith the scope of this thesis and will therefore not be discussed further as part of the analysis.

Data analysis

Exclusion criteria for fMRI participants

Exclusion criteria were specified prior to commencing the fMRI experiment and included that participants' head movements should be no larger than 3mm. One participant was excluded for violating the head movement restrictions.

In order to ensure participants paying attention to the voices during the fMRI task, the hit rate of the scanner responses was calculated for each participant as an average across the four runs. Participants scoring lower than 75% were excluded from the data analysis. A further four participants were removed for violating this criterion. The average hit rate after excluding the total of five participants was 94.5%.

Pre-processing and analysis of fMRI data

Data were pre-processed in Matlab using the standardised SPM8 (Statistical Parametric Mapping, Wellcome Department of Imaging Science, London, UK; <http://www.fil.ion.ucl.ac.uk/spm/software/spm8/>) pipeline. Pre-processing of functional scans consisted of DICOM import/ transformation, ACPC (Anterior Commissure, Posterior Commissure) alignment, slice-time correction to remove any temporal shifts in the scans, correction for head movement, co-registration to the anatomical scans, segmentation into grey matter, white matter, and cerebro-spinal fluid (CSF), and normalisation to MNI space. Normalised data was then smoothed by an 8mm Gaussian kernel at full width half maximum.

Repetition trials (see red voice in Figure 4) were deleted. MNI-normalised bilateral amygdala masks (ANATOMY toolbox within SPM8) for the composite as well as the superficial (SF) amygdala were then used to extract BOLD activation. This was done on each individual voice stimulus for each participant, separately for runs 1, and both runs 1 and 2 combined. Brain activation values were then averaged across the 20 participants, so that each voice stimulus was attributed 8 averaged values of amygdala activation (bilateral: composite run 1; composite runs 1&2; SF run 1; SF runs 1&2). As consensus data appears a better predictor of neural activation (Engell et al., 2007), the averaged amygdala activation values were then correlated (Spearman rank-order correlation) to the behavioural trustworthiness ratings that had been rated in the online validation experiment (Experiment 1), as well as to the averaged group data from the post-scan experiment. This was done separately for female and male voices stimuli.

Additionally, for each ROI it was tested whether a linear or a quadratic relationship were a better fit to the activation pattern in each region. At group level as well as individual level, a linear regression, as well as a second-order polynomial model were fitted to the responses, and a model comparison was performed to investigate whether the two models differed significantly.

Analysis of post-scan behavioural data

Similarly to Experiment 1, participants gave 2 ratings per voice stimuli. The final perceived trustworthiness score for each voice was then calculated as the average of first and second rating the participant gave. Spearman's correlations were calculated for individual participants to investigate how they related to the perceived trustworthiness scores from Experiment 1. For the group level analysis, average trustworthiness ratings from the post-scan behavioural ratings were computed for each voice stimulus, and correlated (Spearman's rank-order correlations) to the average pre-validated scores (Experiment 1). All statistical analysis was conducted in the statistical software R. All critical significance levels were set to .05.

Results

Behavioural results

The average trustworthiness ratings from the post-scan behavioural experiment were correlated to the behavioural ratings obtained in the pre-validation Experiment 1 (see Figure 5). Spearman's rank-order correlation coefficients for female voices between the pre-validated behavioural ratings (Experiment 1), and post-scan behavioural group average were moderate and positive ($\rho = .725$, $p = .002$). For the male voices, Spearman's rank-order correlation calculations revealed significant positive strong correlations ($\rho = .928$, $p < .001$).

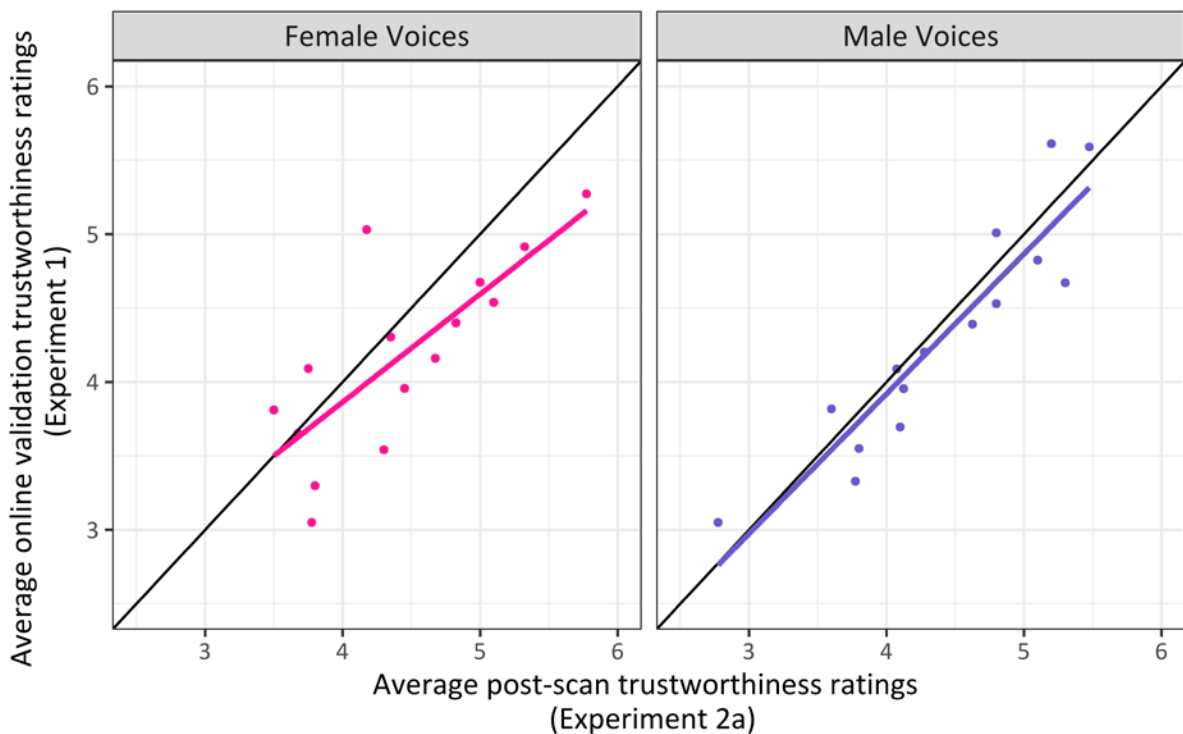


Figure 5: Scatterplot of trustworthiness ratings obtained from the participants in the post-scan behavioural experiment, and in the online validation experiment (linear lines of best fit for male (blue) and female (red) voices, and a reference line (black) to illustrate equal ratings were added)

For individual participants, Spearman's rank-order correlation coefficient were calculated between the post-scan ratings of participants and the ratings obtained in the online validation experiment (Experiment 1). This was done to assess individual differences between participants (Table 2).

Table 2: Spearman's rank-order correlation coefficient for the individual's post-scan ratings and the online validation ratings (Experiment 1), ordered from strongest positive to strongest negative

	Female Voices			Male Voices	
Participants	Spearman's rho	p-value	Participants	Spearman's rho	p-value
Participant 09	.873	<.001	Participant 14	.945	<.001
Participant 19	.816	<.001	Participant 09	.921	<.001
Participant 05	.754	.001	Participant 08	.911	<.001
Participant 08	.750	.001	Participant 13	.900	<.001
Participant 02	.725	.002	Participant 10	.833	<.001
Participant 13	.678	.005	Participant 05	.828	<.001
Participant 04	.640	.010	Participant 02	.791	<.001
Participant 03	.632	.012	Participant 19	.769	.001
Participant 17	.524	.045	Participant 20	.725	.002
Participant 07	.518	.048	Participant 03	.708	.003
Participant 12	.505	.055	Participant 07	.653	.008
Participant 14	.458	.086	Participant 17	.613	.015
Participant 15	.420	.119	Participant 15	.536	.040
Participant 01	.401	.139	Participant 12	.502	.057
Participant 10	.384	.157	Participant 04	.427	.112
Participant 11	.346	.206	Participant 18	.386	.155
Participant 20	.275	.320	Participant 01	.096	.733
Participant 18	.175	.533	Participant 11	-.089	.753
Participant 16	-.051	.857	Participant 16	-.195	.486
Participant 06	-.614	.015	Participant 06	-.523	.045

There is to some extent inter-subject variability of participants' trustworthiness perception in female, and male voices (Table 2). However, 17 of the 20 participants showed significant effects or effects in the direction of the online validation data in the perception of female voices; for male voices there were 16 participants. Two, and three participants showed no directionality for female, and male voices respectively. One participant showed significant effects in the opposite direction to the online validation group, as well as the other

participants for both voice sexes. This could be either due to perceiving trustworthy voices as untrustworthy (and vice versa), or to misunderstanding the rating scale.

Imaging results – female voices

Amygdala activation and online validation behavioural responses

For female voices, Spearman's rank order correlation analysis was employed between the online validation trustworthiness scores (Experiment 1) and the group level amygdala data, utilising the bilateral (L/R) composite (C), and the SF (SF) mask, each for the first run (R1), as well as a combined run 1 and 2 (RC) separately (see Figure 6). No significant Spearman's rank order correlation were found for the left composite amygdala run 1, as well as for combined runs ($\rho_{LCR1} (13) = .089$, $p_{LCR1} = .752$; $\rho_{LCRC} (13) = .386$, $p_{LCRC} = .156$), the right composite amygdala for both run 1 and combined runs ($\rho_{RCR1} (13) = -.146$, $p_{RCR1} = .603$; $\rho_{RCRC} (13) = -.046$, $p_{RCRC} = .869$), the left SF amygdala run 1, and combined runs ($\rho_{LSFR1} (13) = -.036$, $p_{LSFR1} = .899$; $\rho_{LSFRC} (13) = .268$, $p_{LSFRC} = .334$), and the right SF amygdala for both run 1, and combined runs ($\rho_{RSFR1} (13) = -.014$, $p_{RSFR1} = .960$; $\rho_{RSFRC} (13) = .350$, $p_{RSFRC} = .201$).

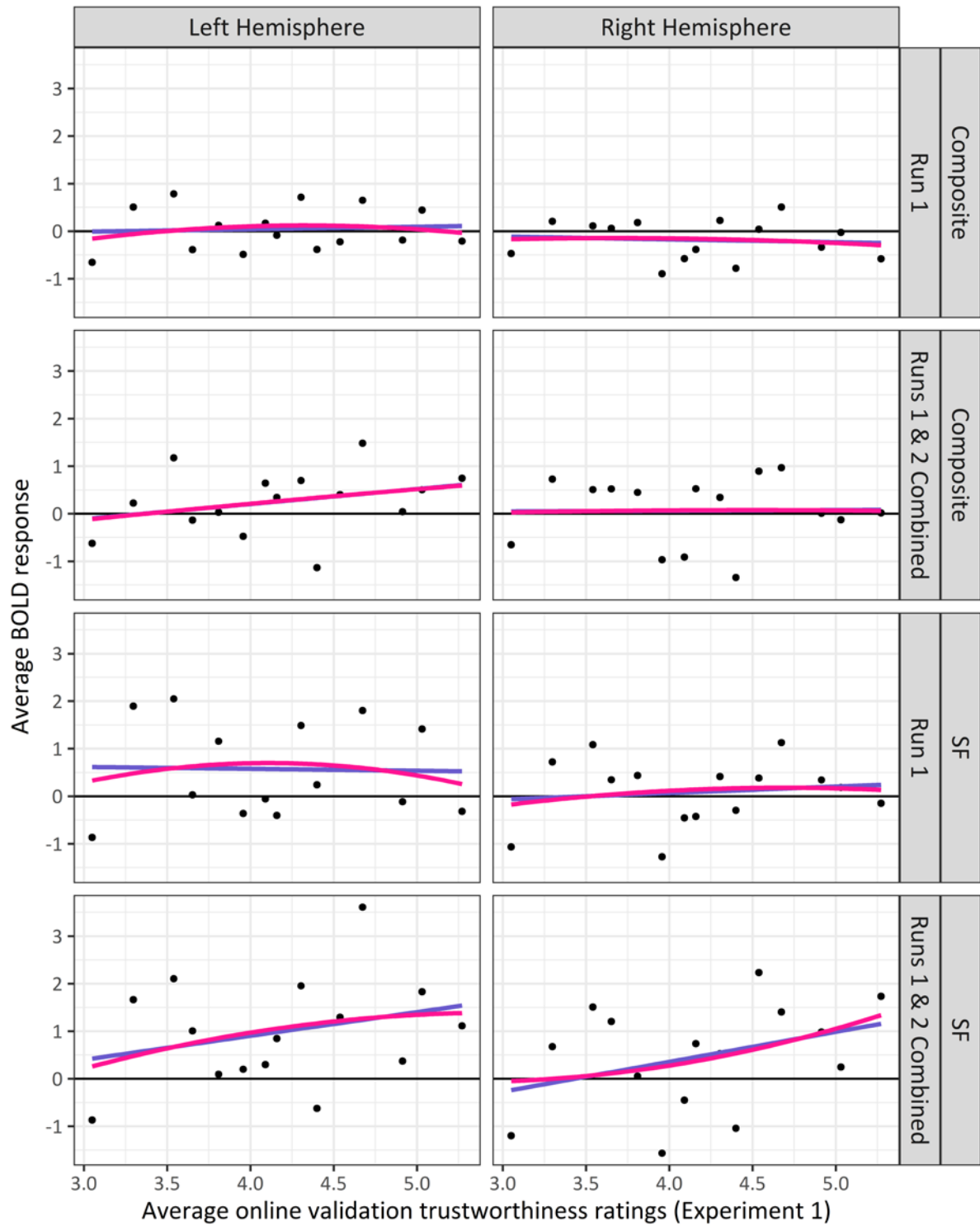


Figure 6: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the female voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

For each of the 8 neural regions displayed in Figure 6, linear and second-order polynomial regression models were fitted for female voices, and a model comparison was subsequently performed to determine which model the better fit is. This was done for the bilateral (L/R) amygdala (composite (C), and superficial (SF) amygdala), separately for the first run (R1), as well as a combined run 1 and 2 (RC). Both, the linear and the quadratic models are not a good fit for the grouped/ averaged data of the female voices, as the models explain only 0-12.92%, and 0-13.62% of the variance in the linear, and quadratic model respectively (Table 3).

Table 3: R² values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and online validation trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R ² linear model	F linear model (p-value)	R ² quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.0050	0.065 (.802)	0.0290	0.179 (.838)	0.296 (.596)
		1&2	0.0898	1.283 (.278)	0.0899	0.593 (.568)	0.001 (.975)
	Right	1	0.0083	0.108 (.747)	0.0119	0.072 (.931)	0.044 (.837)
		1&2	0.0002	0.002 (.964)	0.0003	0.002 (.998)	0.002 (.966)
SF	Left	1	0.0007	0.009 (.925)	0.0215	0.132 (.878)	0.254 (.623)
		1&2	0.0785	1.108 (.312)	0.0835	0.546 (.593)	0.065 (.803)
	Right	1	0.0156	0.206 (.657)	0.0219	0.134 (.876)	0.077 (.787)
		1&2	0.1292	1.929 (.188)	0.1362	0.946 (.416)	0.097 (.761)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R² values for the quadratic polynomials and the linear model, showed the R² of the second-order polynomial to be significantly higher than the R² of the linear model. This held true for the left composite amygdala for both run 1 and combined runs ($t_{LCR1}(19) = -4.686$, $p_{LCR1} < .001$; $t_{LCRC}(19) = -3.831$, $p_{LCRC} = .001$), the right composite amygdala for both run 1 and combined runs ($t_{RCR1}(19) = -4.033$, $p_{RCR1} < .001$; $t_{RCRC}(19) = -3.172$, $p_{RCRC} = .005$), the left SF amygdala for both run 1 and combined runs ($t_{LSFR1}(19) = -3.324$, $p_{LSFR1} = .004$; $t_{LSFRC}(19) = -4.312$, $p_{LSFRC} < .001$), and the right SF amygdala for both run 1 and combined runs ($t_{RSFR1}(19) = -3.156$, $p_{RSFR1} = .005$; $t_{RSFRC}(19) = -5.150$,

$p_{\text{RSFRC}} < .001$). For the individual participant's BOLD response in relation to the online validation data see Supplementary materials (Figure 16, and Figure 17).

Amygdala activation and post-scan behavioural responses

For female voices, Spearman's rank order correlation analysis was employed between the group averages of the post-scan behavioural data and the group level amygdala data, utilising the bilateral (L/R) composite (C), and the SF (SF) mask, each for the first run (R1), as well as a combined run 1 and 2 (RC) separately (see Figure 7). No significant Spearman's rank order correlation were found for the left composite amygdala run 1, as well as for combined runs ($\rho_{\text{LCR1}} (13) = -.061$, $p_{\text{LCR1}} = .830$; $\rho_{\text{LCRC}} (13) = .275$, $p_{\text{LCRC}} = .321$), the right composite amygdala for both run 1 and combined runs ($\rho_{\text{RCR1}} (13) = -.225$, $p_{\text{RCR1}} = .420$; $\rho_{\text{RCRC}} (13) = .089$, $p_{\text{RCRC}} = .752$), the left SF amygdala run 1, and combined runs ($\rho_{\text{LSFR1}} (13) = -.139$, $p_{\text{LSFR1}} = .621$; $\rho_{\text{LSFRC}} (13) = .229$, $p_{\text{LSFRC}} = .413$), and the right SF amygdala for both run 1, and combined runs ($\rho_{\text{RSFR1}} (13) = -.014$, $p_{\text{RSFR1}} = .960$; $\rho_{\text{RSFRC}} (13) = .468$, $p_{\text{RSFRC}} = .079$).

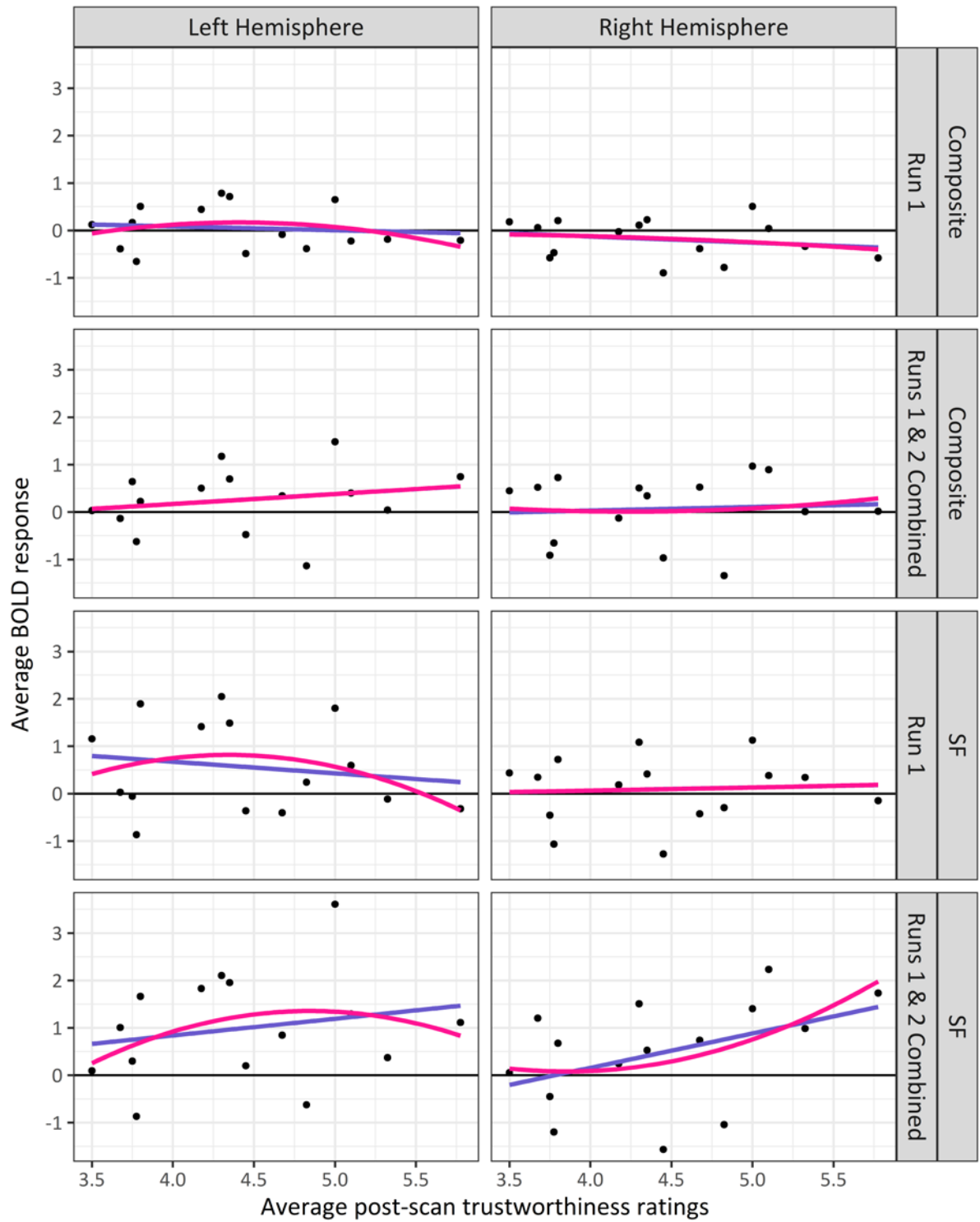


Figure 7: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the female voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

For each of the four ROIs (bilateral (L/R) Composite (C), and SF (SF) amygdala), separately for run 1 and combined runs, linear and second-order polynomial regression models were fitted for female voices. A model comparison was subsequently performed to determine whether the linear or second-order polynomial model the better fit is. This was done for the bilateral (L/R) amygdala (Composite (C), and superficial (SF) amygdala), separately for the first run (R1), as well as a combined run 1 and 2 (RC). Both, the linear and the quadratic models do not explain much variance (0-18.77% in linear models; 0.4-22.85% in quadratic models; Table 4).

Table 4: R² values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and post-scan trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R ² linear model	F linear model (p-value)	R ² quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.0130	0.171 (.686)	0.0802	0.523 (.606)	0.876 (.368)
		1&2	0.0430	0.584 (.458)	0.0430	0.270 (.768)	<0.001 (.995)
	Right	1	0.0462	0.630 (.442)	0.0478	0.301 (.745)	0.020 (.890)
		1&2	0.0049	0.064 (.804)	0.0101	0.061 (.941)	0.062 (.807)
SF	Left	1	0.0285	0.382 (.547)	0.0952	0.631 (.549)	0.883 (.366)
		1&2	0.0432	0.586 (.458)	0.0975	0.648 (.540)	0.722 (.412)
	Right	1	0.0040	0.053 (.822)	0.0040	0.024 (.976)	<0.001 (.994)
		1&2	0.1877	3.005 (.107)	0.2285	1.777 (.211)	0.634 (.442)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R² values for the quadratic polynomials and the linear model, showed the R² of the second-order polynomial to be significantly higher than the R² of the linear model for the left composite amygdala for both run 1 and combined runs ($t_{LCR1}(19) = -3.032$, $p_{LCR1} = .007$; $t_{LCRC}(19) = -3.021$, $p_{LCRC} = .007$), the right composite amygdala for both run 1 and combined runs ($t_{RCR1}(19) = -3.620$, $p_{RCR1} = .002$; $t_{RCRC}(19) = -4.308$, $p_{RCRC} < .001$), the left SF amygdala for both run 1 and combined runs ($t_{LSFR1}(19) = -3.846$, $p_{LSFR1} = .001$; $t_{LSFRC}(19) = -4.052$, $p_{LSFRC} < .001$), and the right SF amygdala for both run 1 and combined runs ($t_{RSFR1}(19) = -3.182$, $p_{RSFR1} = .005$; $t_{RSFRC}(19) = -3.085$, $p_{RSFRC} = .006$). For the

individual participant's BOLD response in relation to the averaged post-scan data see Supplementary materials (Figure 18, and Figure 19).

Imaging results – male voices

Amygdala activation and online validation behavioural responses

For the male voices, Spearman's rank order correlation coefficients were computed between the group averages of the online validation trustworthiness scores (Experiment 1) and the group level amygdala data (C/SF), for the two hemispheres (L/R), and the runs (R1/ RC) separately (see Figure 8). Significant Spearman's rank order correlation were found for the superficial amygdala run 1 bilaterally ($\rho_{LSFR1}(13) = -.525$, $p_{LSFR1} = .044$; $\rho_{RSFR1}(13) = -.525$, $p_{RSFR1} = .044$). After Bonferroni correction for multiple comparisons both p-values are at .356. Correlations for the left composite amygdala ($\rho_{LCR1}(13) = -.443$, $p_{LCR1} = .098$; $\rho_{LCRC}(13) = -.475$, $p_{LCRC} = .074$), right composite amygdala ($\rho_{RCR1}(13) = -.286$, $p_{RCR1} = .302$; $\rho_{RCRC}(13) = -.086$, $p_{RCRC} = .761$), and the bilateral SF for combined runs ($\rho_{LSFRC}(13) = -.454$, $p_{LSFRC} = .089$; $\rho_{RSFRC}(13) = -.350$, $p_{RSFRC} = .201$) were non-significant.

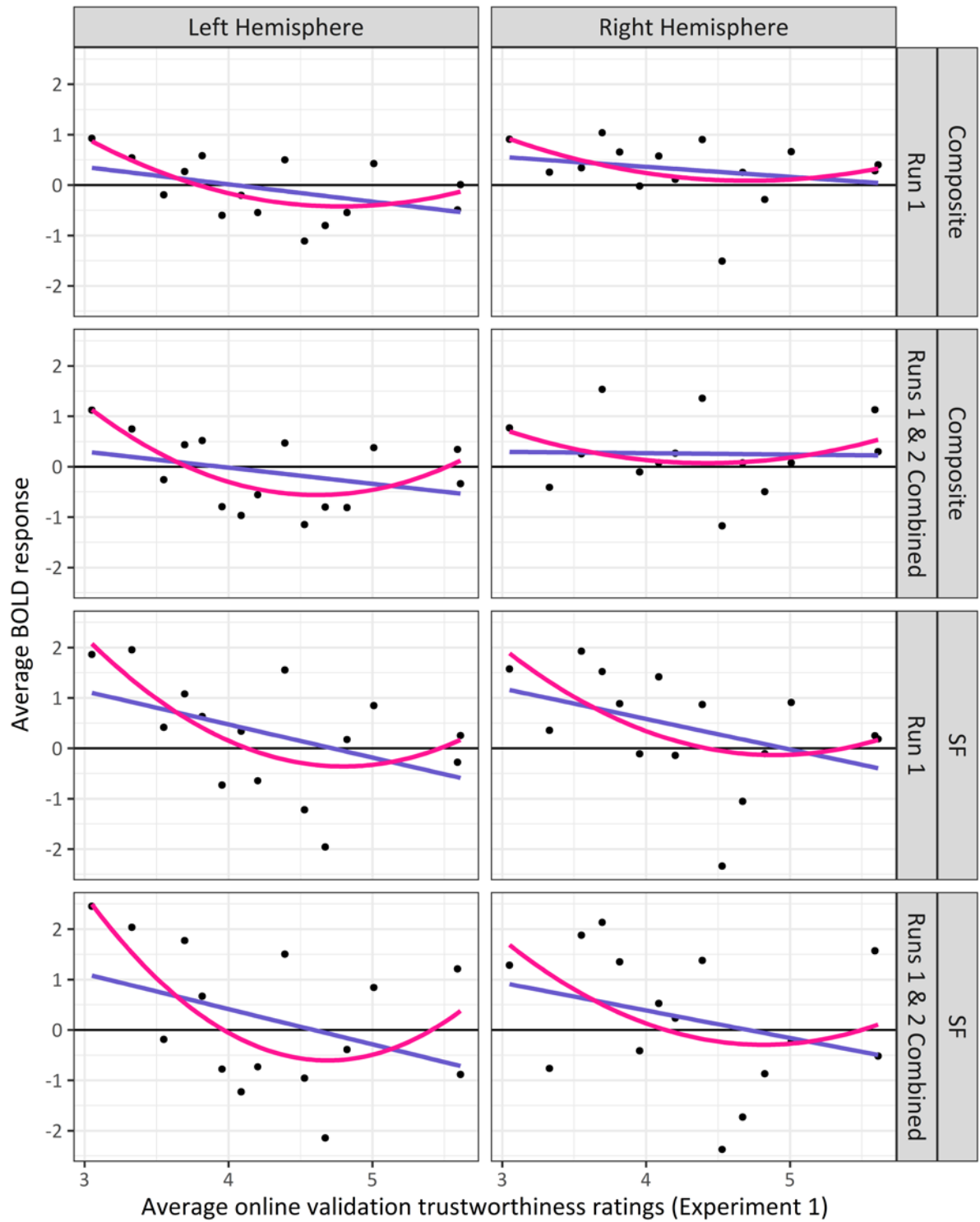


Figure 8: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the male voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

Additionally, a linear and a second level polynomial model were created bilaterally for each ROI in each investigated run (R1/ RC) displayed in Figure 8, and subsequently compared in a model comparison analysis. Both, the linear and the quadratic model are a better fit compared to the models of the female voices. The linear and quadratic model explain respectively 0-20.36% and 8.15-45.68% of the variance of the averaged male trustworthiness ratings. The quadratic models for the left composite and SF amygdala for combined runs are significant ($R^2_{LCRC} = 0.4568$, $F_{LCRC}(2,12) = 5.046$, $p_{LCRC} = .026$; $R^2_{LSFRC} = 0.4112$, $F_{LSFRC}(2,12) = 4.190$, $p_{LSFRC} = .042$), and significantly better than the linear models ($F_{LCRC} = 7.556$, $p_{LCRC} = .018$; $F_{LSFRC} = 5.303$, $p_{LSFRC} = .040$). The quadratic model for the left SF amygdala run 1 is significant ($R^2_{LSFR1} = 0.3929$, $F_{LSFR1}(2,12) = 3.883$, $p_{LSFR1} = .050$), albeit not significantly different to the linear model ($F_{FSFR1} = 3.741$, $p_{LSFR1} = .077$). All other model comparisons reached non-significant results (Table 5).

Table 5: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and online validation trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R^2 linear model	F linear model (p-value)	R^2 quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.1913	3.075 (.103)	0.3837	3.736 (.055)	3.747 (.077)
		1&2	0.1148	1.686 (.217)	0.4568	5.046 (.026)	7.556 (.018)
	Right	1	0.0594	0.821 (.381)	0.1459	1.025 (.388)	1.216 (.292)
		1&2	0.0009	0.011 (.917)	0.0815	0.532 (.601)	1.053 (.325)
SF	Left	1	0.2036	3.324 (.091)	0.3929	3.883 (.050)	3.741 (.077)
		1&2	0.1509	2.311 (.152)	0.4112	4.190 (.042)	5.303 (.040)
	Right	1	0.1753	2.763 (.120)	0.2812	2.348 (.138)	1.769 (.208)
		1&2	0.0951	1.366 (.263)	0.1761	1.282 (.313)	1.180 (.299)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R^2 values for the quadratic polynomials and the linear model, was computed. The t-test showed that R^2 of the second-order polynomial was significantly higher than the R^2 of the linear model in both run 1 and combined runs for the left composite amygdala ($t_{LCR1}(19) = -3.100$, $p_{LCR1} = .006$; $t_{LCRC}(19) = -3.630$, $p_{LCRC} = .002$), the

right composite amygdala ($t_{RCR1}(19) = -3.919$, $p_{RCR1} < .001$; $t_{RCRC}(19) = -2.856$, $p_{RCRC} = .010$), the left SF amygdala ($t_{LSFR1}(19) = -2.900$, $p_{LSFR1} = .009$; $t_{LSFRC}(19) = -4.521$, $p_{LSFRC} < .001$), and the right SF amygdala ($t_{RSFR1}(19) = -3.497$, $p_{RSFR1} = .002$; $t_{RSFRC}(19) = -3.643$, $p_{RSFRC} = .002$). For the individual participant's BOLD response in relation to the online validation data see Supplementary materials (Figure 20, and Figure 21).

Amygdala activation and post-scan behavioural responses

For the male voices, Spearman's rank order correlation coefficients were computed between the group averages of the post-scan behavioural responses and the group level amygdala data (C/SF), for the two hemispheres (L/R), and the runs (R1/ RC) separately (see Figure 9). Significant Spearman's rank order correlation were found for the left composite amygdala run 1 and combined runs ($\rho_{LCR1}(13) = -.604$, $p_{LCR1} = .017$; $\rho_{LCRC}(13) = -.517$, $p_{LCRC} = .049$), the bilateral superficial amygdala run 1 ($\rho_{LSFR1}(13) = -.624$, $p_{LSFR1} = .013$; $\rho_{RSFR1}(13) = -.622$, $p_{RSFR1} = .013$). After Bonferroni correction for multiple comparisons all p-values would be non-significant ($p_{LCR1} = .136$; $p_{LCRC} = .392$; $p_{LSFR1} = .104$; $p_{RSFR1} = .104$). Correlations for the right composite amygdala for both run 1 and combined runs ($\rho_{RCR1}(13) = -.399$, $p_{RCR1} = .141$; $\rho_{RCRC}(13) = -.068$, $p_{RCRC} = .810$), and bilateral SF amygdala for combined runs ($\rho_{LSFRC}(13) = -.438$, $p_{LSFRC} = .103$; $\rho_{RSFRC}(13) = -.368$, $p_{RSFRC} = .177$) were non-significant.

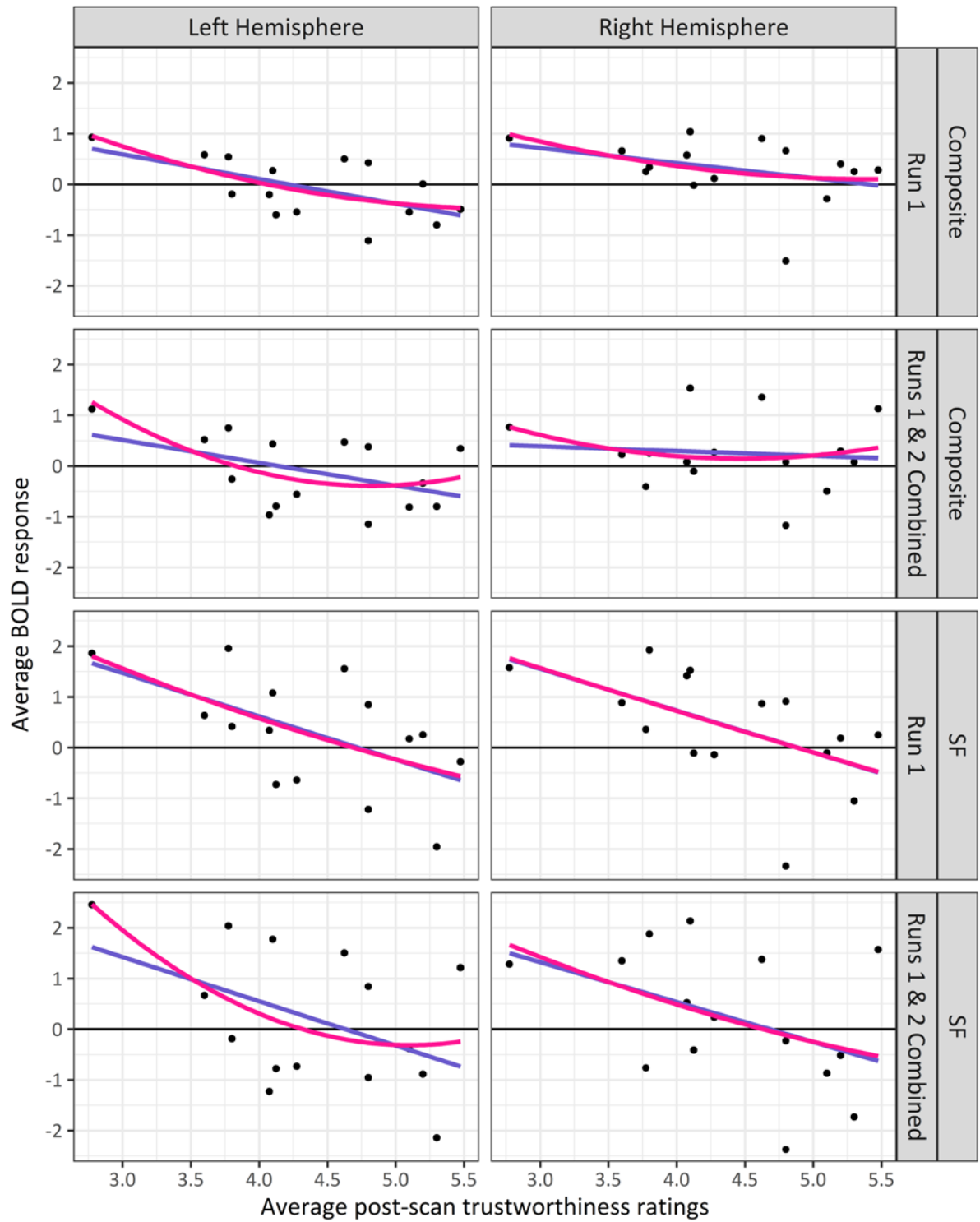


Figure 9: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the male voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

Additionally, a linear and a second level polynomial model were created bilaterally for each ROI in each investigated run (R1/ RC) displayed in Figure 9, and subsequently compared in a model comparison analysis. Similarly to the models created with the online validation data, both, the linear and the quadratic models are a better fit than the models regarding the female voices. In the linear models 0.96-36.37% of the variance are explained whereas the quadratic models explains 4.67-39.05%. The linear models of the first runs for the left composite amygdala ($R^2_{LCR1} = 0.3637$, $F_{LCR1}(1,13) = 7.431$, $p_{LCR1} = .017$), and the bilateral SF amygdala ($R^2_{LSFR1} = 0.3244$, $F_{LSFR1}(1,13) = 6.243$, $p_{LSFR1} = .027$; $R^2_{RSFR1} = 0.3083$, $F_{RSFR1}(1,13) = 5.795$, $p_{RSFR1} = .032$) are significant, though all model comparisons were non-significant (Table 6).

Table 6: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and post-scan trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R^2 linear model	F linear model (p-value)	R^2 quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.3637	7.431 (.017)	0.3905	3.844 (.051)	0.527 (.482)
		1&2	0.2161	3.583 (.081)	0.3368	3.046 (.085)	2.184 (.165)
	Right	1	0.1275	1.901 (.191)	0.1437	1.007 (.394)	0.226 (.643)
		1&2	0.0096	0.126 (.728)	0.0467	0.294 (.751)	0.466 (.508)
SF	Left	1	0.3244	6.243 (.027)	0.3267	2.912 (.093)	0.041 (.842)
		1&2	0.2221	3.711 (.076)	0.2781	2.311 (.142)	0.930 (.354)
	Right	1	0.3083	5.795 (.032)	0.3084	2.675 (.109)	0.001 (.978)
		1&2	0.1862	2.975 (.108)	0.1884	1.393 (.286)	0.032 (.862)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R^2 values for the quadratic polynomials and the linear model, showed the R^2 of the second-order polynomial to be significantly higher than the R^2 of the linear model. This held true for the left composite amygdala for both run 1 and combined runs ($t_{LCR1}(19) = -3.185$, $p_{LCR1} = .005$; $t_{LCRC}(19) = -4.740$, $p_{LCRC} < .001$), the right composite amygdala for both run 1 and combined runs ($t_{RCR1}(19) = -3.166$, $p_{RCR1} = .005$; $t_{RCRC}(19) = -2.963$, $p_{RCRC} = .008$), the left SF amygdala for both run 1 and combined runs

($t_{\text{LSFR1}}(19) = -3.993$, $p_{\text{LSFR1}} < .001$; $t_{\text{LSFRC}}(19) = -4.395$, $p_{\text{LSFRC}} < .001$), and the right SF amygdala for both run 1 and combined runs ($t_{\text{RSFR1}}(19) = -3.844$, $p_{\text{RSFR1}} = .001$; $t_{\text{RSFRC}}(19) = -4.935$, $p_{\text{RSFRC}} < .001$). For the individual participant's BOLD response in relation to the online validation data see Supplementary materials (Figure 22, and Figure 23).

Summary of results

The aim of Experiment 2a was to investigate the relationship between amygdala activation and perceptions of trustworthiness with an implicit 1-back task design high in attention and cognitive load. Behaviourally, perceived trustworthiness ratings, gathered in the post-scan experiment, correlated highly with the consensus data obtained in the online validation experiment (Experiment 1) suggesting that both sets of participants perceived the trustworthy voices as trustworthy, and the untrustworthy voices as untrustworthy. This held true for female as well as male voice stimuli.

Amygdala activation was then correlated to the trustworthiness ratings (Experiment 1), and significant negative moderate correlations were found for the bilateral SF amygdala (run 1) for male but not female voices. In a validation attempt, amygdala activation was then correlated to the group averages of the post-scan trustworthiness ratings, and correlation values were again negative and moderate for male voice stimuli, however, regions involved were the bilateral composite amygdala for run 1, and the left SF amygdala for both runs. No significant correlations can be reported for female voices. Bonferroni corrections had to be applied given the exploratory nature of the experiment, leaving all p-values non-significant. Despite, this shows that the amygdala is sensitive to varying levels of perceived vocal trustworthiness but this needs to be replicated in a follow-up confirmatory experiment. Furthermore, a linear and quadratic model comparison on the individual data showed that quadratic models are a better predictor of amygdala activation, which is in agreement with findings from face research (i.e. Mattavelli et al., 2012). Further discussion of the results in light of theory will be held for the General discussion and conclusion (page 69) taking findings from Experiment 2b into consideration.

Experiment 2b: fMRI experiment – PureTone detection task

Experiment 2b aimed to investigate correlations between amygdala activation and varying levels of vocal trustworthiness. Equivalent to Experiment 2a, the bilateral SF and the bilateral composite amygdala were selected as ROIs for this study as those regions have been identified in relation to facial trustworthiness (Bzdok et al., 2011; Mende-Siedlecki et al., 2013; Todorov & Engell, 2008). Thirty voice stimuli (15 females) saying the word ‘hello’ were selected during online validation in Experiment 1. Similarly to Experiment 2a, an implicit design was employed as no instructions in regards to perceived trustworthiness were provided to participants. Here a PureTone detection task was created on the basis of the red spot detection task used by Mattavelli et al. (2012) in which participants identify whether the stimulus played is a beep tone or a voice. Compared to the 1-back task in Experiment 2a, the cognitive load of the PureTone detection task is reduced. Equivalent to Experiment 2a, behavioural responses of how trustworthy participants perceived the voices to be, were gathered subsequent to the fMRI experiment. Again, age was kept consistent (between 17 and 30 years), and voice sexes were analysed separately. As in Experiment 2a, amygdala activation was correlated to the consensus data of Experiment 1, and the post-scan behavioural results. A secondary aim of this study was to explore whether a linear or quadratic response pattern was a better fit for the neural data (see Bzdok et al., 2011; Freeman et al., 2014; Mattavelli et al., 2012).

Methods

Unless stated otherwise, the methods were the same as in Experiment 2a. Please refer to the previous chapter.

Participants

Advertising and recruitment criteria, and monetary incentives for partaking were the same as in Experiment 2a. Twenty participants (10 male, average age: 21.25 ± 2.53 , age range: 17-26) took part in the fMRI experiment. An additional two participated but were excluded during data analysis (see Exclusion criteria for fMRI participants, page 51).

fMRI paradigm, and procedure

fMRI paradigm, and procedure were the same as in Experiment 2a except that participants completed a PureTone detection task instead of the 1-back task. They were instructed to pay attention to the voices, and press a button when they hear a beep rather than a repeated voice. Each block consisted of 15 voice stimuli and either one, or two beeps (Figure 10). Similarly to experiment 2a, the order of the runs, the position of the beeps, and the order of the voice stimuli within each block was counterbalanced.

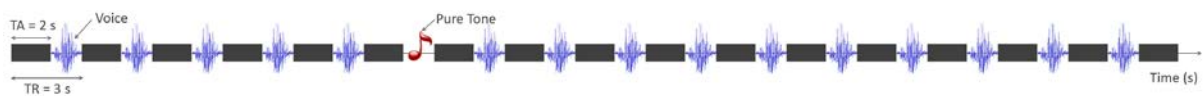


Figure 10: Diagram depicting one block of Sparse Sampling paradigm with 15 voice stimuli and one beep (adapted from Bestelmeyer et al., 2012)

Data analysis

Exclusion criteria for fMRI participants

As in Experiment 2a, the head movement exclusion criterion was set to 3mm. One participant was excluded.

The hit rate criterion of the scanner responses was increased to a 90% threshold due to decreasing task demands. One participant scored 81% and was hence excluded from the data analysis. The average hit rate after excluding the total of five participants was 99.5%.

Pre-processing and analysis of fMRI data

Pre-processing and analysis of fMRI data were equivalent to Experiment 2a except that PureTone trials (see red musical note in Figure 10) instead of the repetition trials being removed before applying the amygdala masks for extracting BOLD activation in the ROIs.

Results

Behavioural results

Spearman's rank-order correlation coefficients were calculated between the average trustworthiness ratings from all participants in the post-scan behavioural experiment and the behavioural ratings obtained in the pre-validation Experiment 1 (Figure 11). For both the female and male voices, Spearman's rho was revealed to be strong and positive ($\rho_{\text{FemaleVoices}} = .839, p < .001$; $\rho_{\text{MaleVoices}} = .958, p < .001$).

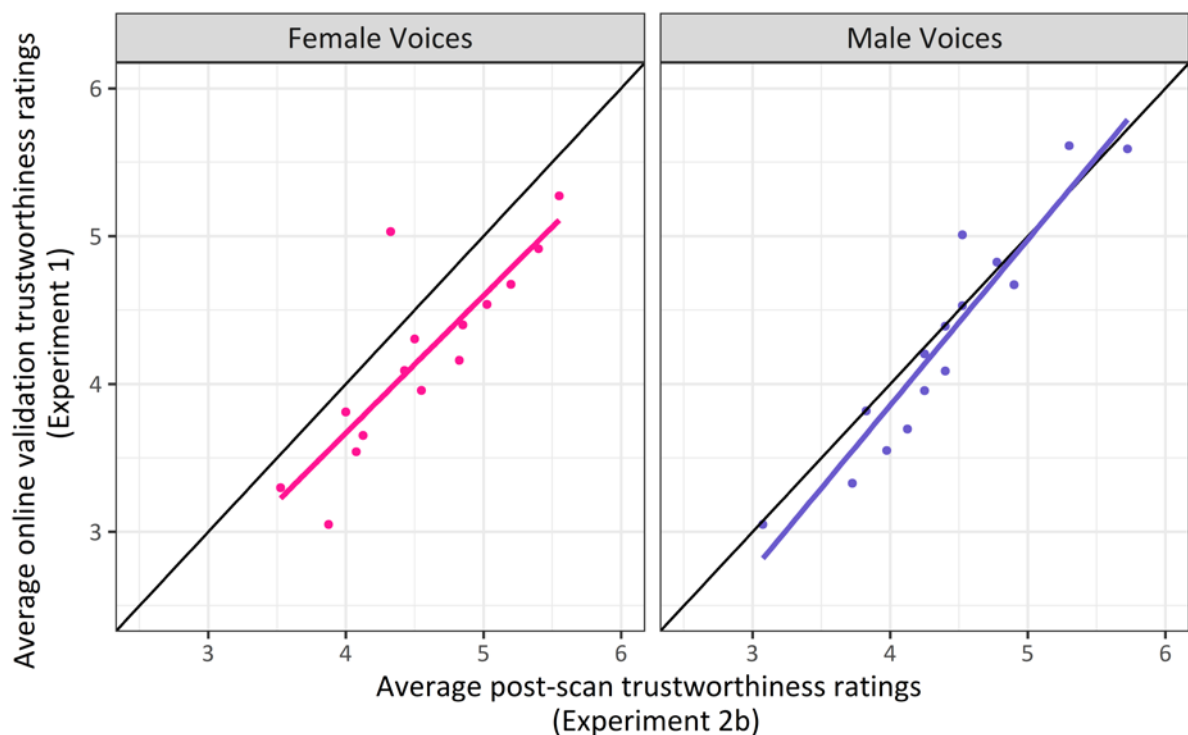


Figure 11: Scatterplot of trustworthiness ratings obtained from the participants in the post-scan behavioural experiment, and in the online validation experiment (linear lines of best fit for male (blue) and female (red) voices, and a reference line (black) to illustrate equal ratings were added)

For individual participants, Spearman's rank-order correlation coefficient were calculated between participants' post-scan ratings and the pre-validated online trustworthiness ratings (Experiment 1). Similarly to the participants in Experiment 2a, there is inter-subject variety within participants' perception of trustworthiness for the female, and male voices (Table 7). Fifteen out of 20 participants showed significant effects or effects in the direction of the online validation data in the perception of trustworthiness in female and male voices. There

are four and five participants who displayed no directionality for female or male voices respectively. One participant showed effects in the opposite direction for female voices, albeit non-significant.

Table 7: Spearman's rank-order correlation coefficient for the individual's post-scan ratings and the online validation ratings (Experiment 1), ordered from strongest positive to strongest negative

	Female Voices			Male Voices	
Participants	Spearman's rho	p value	Participants	Spearman's rho	p value
Participant 40	.918	<.001	Participant 24	.898	<.001
Participant 36	.858	<.001	Participant 35	.880	<.001
Participant 39	.837	<.001	Participant 33	.866	<.001
Participant 24	.836	<.001	Participant 21	.849	<.001
Participant 29	.820	<.001	Participant 39	.826	<.001
Participant 35	.705	.003	Participant 27	.820	<.001
Participant 30	.694	.004	Participant 34	.768	.001
Participant 33	.617	.014	Participant 36	.681	.005
Participant 27	.559	.030	Participant 30	.669	.006
Participant 37	.469	.078	Participant 32	.649	.009
Participant 38	.403	.137	Participant 37	.606	.017
Participant 32	.402	.138	Participant 40	.331	.228
Participant 26	.370	.175	Participant 29	.322	.243
Participant 21	.334	.223	Participant 22	.311	.258
Participant 25	.331	.229	Participant 25	.223	.425
Participant 22	.183	.515	Participant 28	.179	.523
Participant 28	.129	.646	Participant 26	.177	.529
Participant 34	-.030	.916	Participant 31	.170	.544
Participant 31	-.088	.756	Participant 23	.065	.818
Participant 23	-.238	.392	Participant 38	-.095	.738

Imaging results – female voices

Amygdala activation and online validation behavioural responses

For female voices Spearman's rank order correlation analysis was employed between the online validation trustworthiness scores (Experiment 1) and the group level amygdala data, utilising the bilateral (L/R) Composite (C), and the SF (SF) mask, each for the first run (R1), as well as a combined run 1 and 2 (RC) separately (see Figure 12). No significant Spearman's rank order correlation were found for the left composite amygdala run 1, as well as for combined runs ($\rho_{LCR1}(13) = .154$, $p_{LCR1} = .585$; $\rho_{LCRC}(13) = -.061$, $p_{LCRC} = .830$), the right composite amygdala for both run 1 and combined runs ($\rho_{RCR1}(13) = -.211$, $p_{RCR1} = .451$; $\rho_{RCRC}(13) = -.014$, $p_{RCRC} = .960$), the left SF amygdala run 1, and combined runs ($\rho_{LSFR1}(13) = .150$, $p_{LSFR1} = .594$; $\rho_{LSFRC}(13) = .004$, $p_{LSFRC} = .990$), and the right SF amygdala for both run 1, and combined runs ($\rho_{RSFR1}(13) = -.264$, $p_{RSFR1} = .341$; $\rho_{RSFRC}(13) = -.111$, $p_{RSFRC} = .694$).

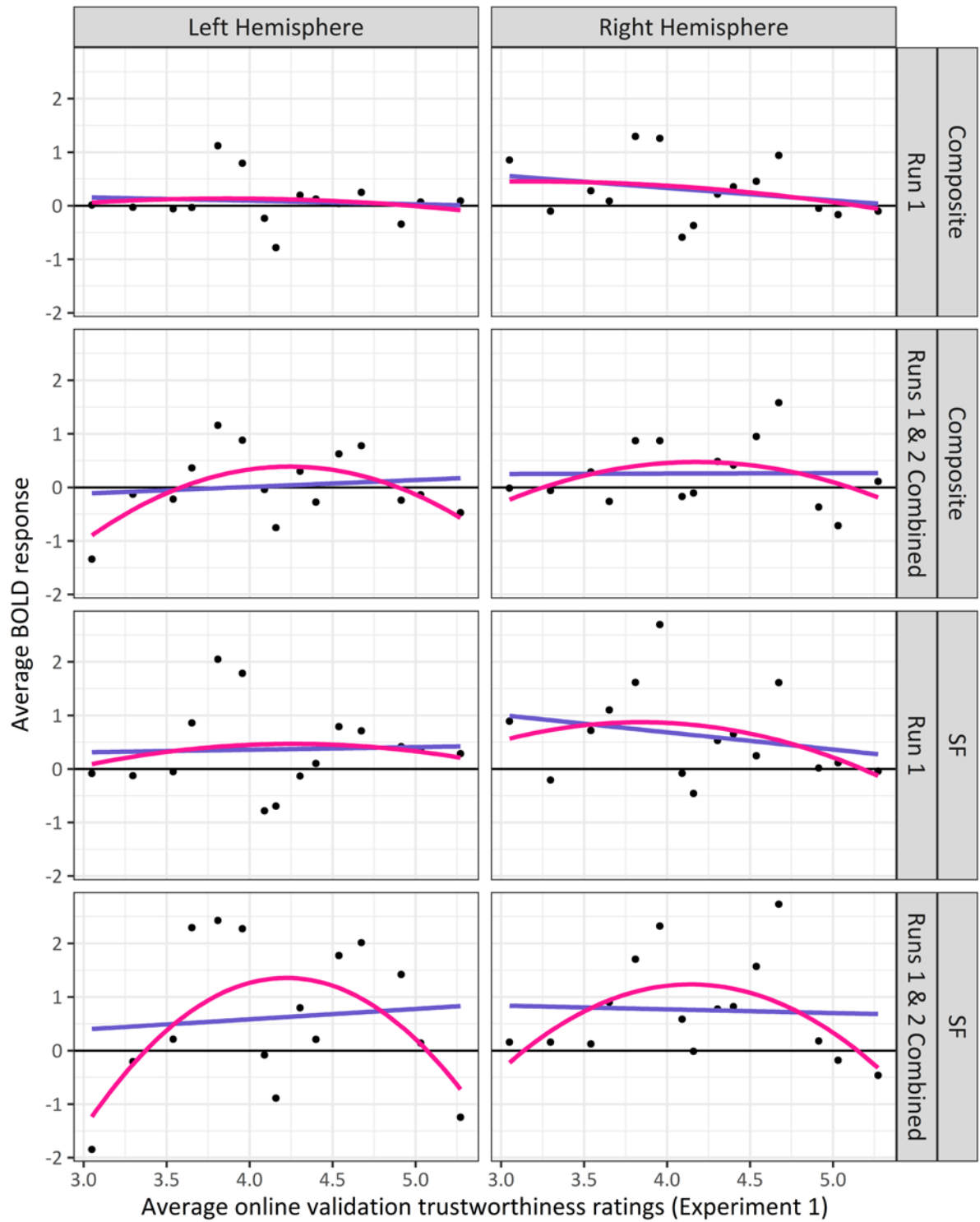


Figure 12: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the female voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

Linear and second-order polynomial regression models were fitted to the averaged group data for female voices, and a model comparison was subsequently performed to determine which model the better fit is. This was done for the bilateral (L/R) Composite (C), and the SF (SF) amygdala, separately for the first run (R1), as well as a combined run 1 and 2 (RC). The linear models are not a good fit for the grouped/ averaged data of the female voices as only 0-6.78% of the variance are explained. The quadratic models are a slightly better fit, explaining 2.04-35.71% of the variance, albeit all differences between the two models were non-significant (Table 8).

Table 8: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and online validation trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R^2 linear model	F linear model (p-value)	R^2 quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.0096	0.126 (.729)	0.0211	0.129 (.880)	0.141 (.714)
		1&2	0.0156	0.207 (.657)	0.3571	3.332 (.071)	6.373 (.027)
	Right	1	0.0678	0.946 (.349)	0.0754	0.489 (.625)	0.098 (.760)
		1&2	0.0001	0.001 (.978)	0.1510	1.067 (.375)	2.133 (.170)
SF	Left	1	0.0016	0.020 (.888)	0.0204	0.125 (.884)	0.231 (.640)
		1&2	0.0082	0.108 (.748)	0.3536	3.282 (.073)	6.411 (.026)
	Right	1	0.0597	0.826 (.380)	0.1200	0.818 (.465)	0.821 (.383)
		1&2	0.0023	0.029 (.866)	0.3101	2.697 (.108)	5.354 (.039)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R^2 values for the quadratic polynomials and the linear model, showed the R^2 of the second-order polynomial to be significantly higher than the R^2 of the linear model for the left composite amygdala for both run 1 and combined runs ($t_{LCR1}(19) = -3.681$, $p_{LCR1} = .002$; $t_{LCRC}(19) = -4.049$, $p_{LCRC} < .001$), the right composite amygdala for both run 1 and combined runs ($t_{RCR1}(19) = -2.681$, $p_{RCR1} = .015$; $t_{RCRC}(19) = -4.716$, $p_{RCRC} < .001$), the left SF amygdala for both run 1 and combined runs ($t_{LSFR1}(19) = -4.740$, $p_{LSFR1} < .001$; $t_{LSFRC}(19) = -4.284$, $p_{LSFRC} < .001$), and the right SF amygdala for both run 1 and combined runs ($t_{RSFR1}(19) = -3.501$, $p_{RSFR1} = .002$; $t_{RSFRC}(19) = -3.964$, $p_{RSFRC} < .001$). For the

individual participant's BOLD response in relation to the online validation data see Supplementary materials (Figure 24, and Figure 25).

Amygdala activation and post-scan behavioural responses

For female voices Spearman's rank order correlation analysis was employed between the group averages of the post-scan behavioural data and the group level amygdala data, utilising the bilateral (L/R) Composite (C), and the SF (SF) mask, each for the first run (R1), as well as a combined run 1 and 2 (RC) separately (see Figure 13). No significant Spearman's rank order correlation were found for female voices for the left composite amygdala run 1, as well as for combined runs ($\rho_{LCR1}(13) = .061$, $p_{LCR1} = .830$; $\rho_{LCRC}(13) = -.082$, $p_{LCRC} = .771$), the right composite amygdala for both run 1 and combined runs ($\rho_{RCR1}(13) = -.039$, $p_{RCR1} = .889$; $\rho_{RCRC}(13) = .182$, $p_{RCRC} = .516$), the left SF amygdala run 1, and combined runs ($\rho_{LSFR1}(13) = .154$, $p_{LSFR1} = .585$; $\rho_{LSFRC}(13) = .064$, $p_{LSFRC} = .820$), and the right SF amygdala for both run 1, and combined runs ($\rho_{RSFR1}(13) = -.164$, $p_{RSFR1} = .558$; $\rho_{RSFRC}(13) = .086$, $p_{RSFRC} = .761$).

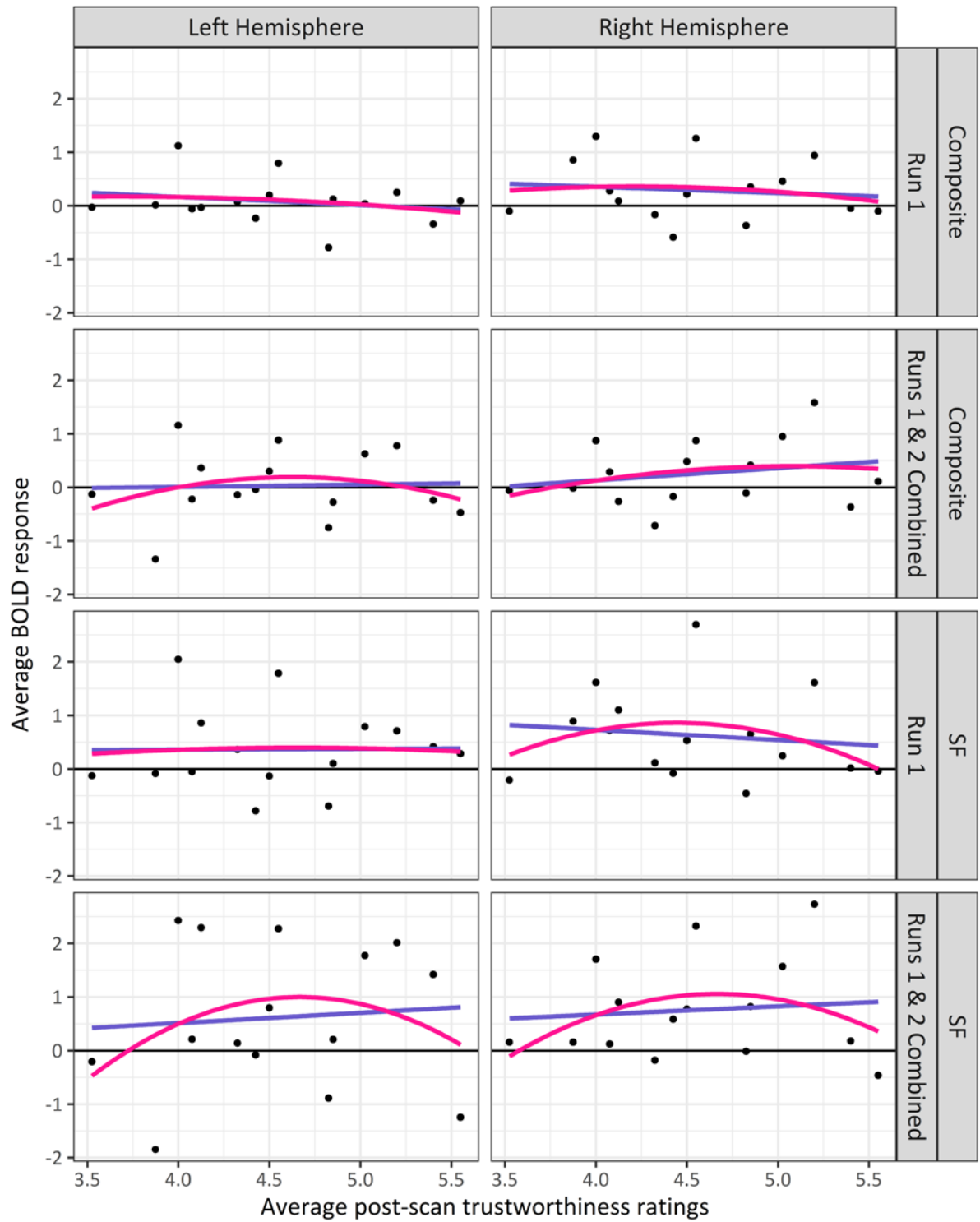


Figure 13: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the female voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

Linear and second-order polynomial regression models were fitted to the averaged group data for female voices, and a model comparison was subsequently performed to determine which model the better fit is. This was done for the bilateral (L/R) Composite (C), and the SF (SF) amygdala, separately for the first run (R1), as well as a combined run 1 and 2 (RC). Both, the linear and the quadratic models are not a good fit for the grouped/ averaged data of the female voices as only 0-4.81% and 0.16-11.79% of the data variance are explained by them respectively (Table 9).

Table 9: R² values for the linear model, the quadratic polynomial model, and a model comparison for female voice sex and post-scan trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R ² linear model	F linear model (p-value)	R ² quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.0425	0.577 (.461)	0.0470	0.296 (.749)	0.056 (.816)
		1&2	0.0015	0.019 (.892)	0.0661	0.425 (.663)	0.831 (.380)
	Right	1	0.0141	0.186 (.673)	0.0233	0.143 (.868)	0.112 (.743)
		1&2	0.0481	0.657 (.432)	0.0648	0.416 (.669)	0.214 (.652)
SF	Left	1	0.0001	0.001 (.973)	0.0016	0.010 (.991)	0.018 (.896)
		1&2	0.0066	0.086 (.773)	0.0889	0.586 (.572)	1.084 (.318)
	Right	1	0.0173	0.229 (.641)	0.1001	0.667 (.531)	1.104 (.314)
		1&2	0.0091	0.119 (.735)	0.1179	0.802 (.471)	1.480 (.247)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R² values for the quadratic polynomials and the linear model, showed the R² of the second-order polynomial to be significantly higher than the R² of the linear model for the left composite amygdala for both run 1 and combined runs ($t_{LCR1}(19) = -4.013$, $p_{LCR1} < .001$; $t_{LCRC}(19) = -4.586$, $p_{LCRC} < .001$), the right composite amygdala for both run 1 and combined runs ($t_{RCR1}(19) = -4.147$, $p_{RCR1} < .001$; $t_{RCRC}(19) = -5.154$, $p_{RCRC} < .001$), the left SF amygdala for both run 1 and combined runs ($t_{LSFR1}(19) = -5.558$, $p_{LSFR1} < .001$; $t_{LSFRC}(19) = -4.818$, $p_{LSFRC} < .001$), and the right SF amygdala for both run 1 and combined runs ($t_{RSFR1}(19) = -5.132$, $p_{RSFR1} < .001$; $t_{RSFRC}(19) = -4.375$, $p_{RSFRC} < .001$). For the

individual participant's BOLD response in relation to the online validation data see Supplementary materials (Figure 26, Figure 27).

Imaging results – male voices

Amygdala activation and online validation behavioural responses

For the male voices, Spearman's rank order correlation coefficients were computed between the online validation trustworthiness scores (Experiment 1) and the group level amygdala data (C/SF), for the two hemispheres (L/R), and the runs (R1/ RC) separately (see Figure 14). No significant Spearman's rank order correlation were found for the left composite amygdala run 1, as well as for combined runs ($\rho_{LCR1}(13) = .379$, $p_{LCR1} = .164$; $\rho_{LCRC}(13) = .368$, $p_{LCRC} = .177$), the right composite amygdala for both run 1 and combined runs ($\rho_{RCR1}(13) = .339$, $p_{RCR1} = .216$; $\rho_{RCRC}(13) = .314$, $p_{RCRC} = .254$), the left SF amygdala run 1, and combined runs ($\rho_{LSFR1}(13) = .361$, $p_{LSFR1} = .187$; $\rho_{LSFRC}(13) = .168$, $p_{LSFRC} = .550$), and the right SF amygdala for both run 1, and combined runs ($\rho_{RSFR1}(13) = .318$, $p_{RSFR1} = .248$; $\rho_{RSFRC}(13) = .225$, $p_{RSFRC} = .420$).

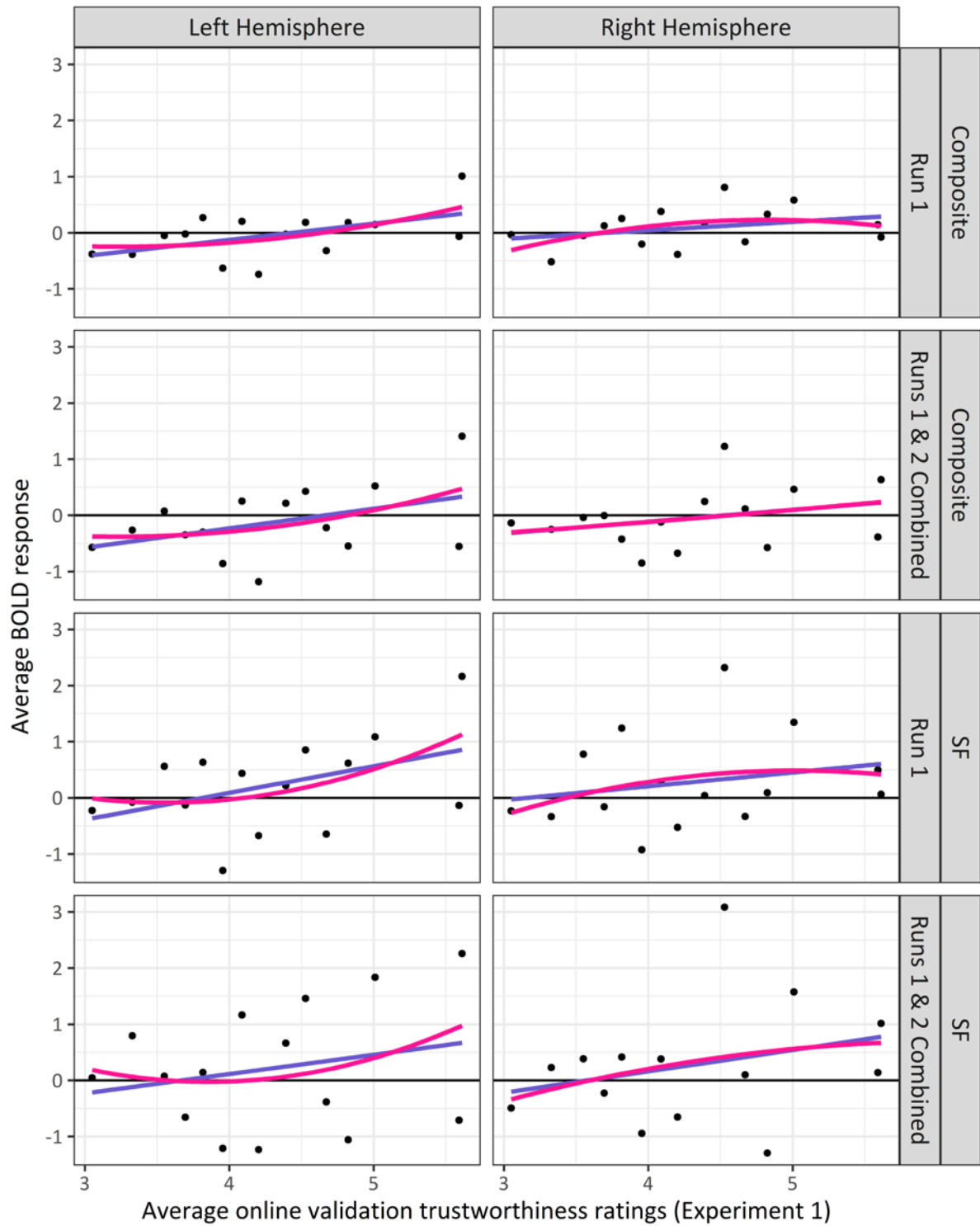


Figure 14: Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1) of the male voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

Additionally, a linear and a second level polynomial model were created bilaterally for each ROI in each investigated run (R1/ RC), and subsequently compared in a model comparison analysis. Both, the linear and the quadratic models explain more of the averaged male trustworthiness ratings than they did for the female voices. In the linear model 5-26.9% of the variance are explained whereas in the quadratic models R^2 was 6.97-30.07%. Though, all of the model comparisons for the male voices were non-significant (Table 10).

Table 10: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and online validation trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R^2 linear model	F linear model (p-value)	R^2 quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.2690	4.783 (.048)	0.3007	2.580 (.117)	0.544 (.475)
		1&2	0.1744	2.747 (.121)	0.1952	1.455 (.272)	0.309 (.589)
	Right	1	0.1070	1.558 (.234)	0.1958	1.460 (.271)	1.324 (.272)
		1&2	0.0890	1.270 (.280)	0.0891	0.587 (.571)	0.002 (.968)
SF	Left	1	0.1897	3.043 (.105)	0.2337	1.830 (.202)	0.690 (.422)
		1&2	0.0553	0.761 (.399)	0.0866	0.569 (.581)	0.411 (.533)
	Right	1	0.0500	0.684 (.423)	0.0697	0.450 (.648)	0.255 (.623)
		1&2	0.0749	1.053 (.324)	0.0789	0.514 (.611)	0.052 (.823)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R^2 values for the quadratic polynomials and the linear model, showed the R^2 of the second-order polynomial to be significantly higher than the R^2 of the linear model for the left composite amygdala for both run 1 and combined runs ($t_{LCR1}(19) = -4.064$, $p_{LCR1} < .001$; $t_{LCRC}(19) = -4.577$, $p_{LCRC} < .001$), the right composite amygdala for both run 1 and combined runs ($t_{RCR1}(19) = -3.222$, $p_{RCR1} = .005$; $t_{RCRC}(19) = -2.932$, $p_{RCRC} = .009$), the left SF amygdala for both run 1 and combined runs ($t_{LSFR1}(19) = -4.070$, $p_{LSFR1} < .001$; $t_{LSFRC}(19) = -4.351$, $p_{LSFRC} < .001$), and the right SF amygdala for both run 1 and combined runs ($t_{RSFR1}(19) = -3.644$, $p_{RSFR1} = .002$; $t_{RSFRC}(19) = -2.916$, $p_{RSFRC} = .009$). For the individual participant's BOLD response in relation to the online validation data see Supplementary materials (Figure 28, and Figure 29).

Amygdala activation and post-scan behavioural responses

For the male voices, Spearman's rank order correlation coefficients were computed between the group averages of the post-scan behavioural data for the PureTone participants and the group level amygdala data (C/SF), for the two hemispheres (L/R), and the runs (R1/ RC) separately (see Figure 14). No significant Spearman's rank order correlation were found for the left composite amygdala run 1, as well as for combined runs ($\rho_{LCR1}(13) = .326$, $p_{LCR1} = .236$; $\rho_{LCRC}(13) = .342$, $p_{LCRC} = .212$), the right composite amygdala for both run 1 and combined runs ($\rho_{RCR1}(13) = .308$, $p_{RCR1} = .264$; $\rho_{RCRC}(13) = .324$, $p_{RCRC} = .239$), the left SF amygdala run 1, and combined runs ($\rho_{LSFR1}(13) = .242$, $p_{LSFR1} = .388$; $\rho_{LSFRC}(13) = .097$, $p_{LSFRC} = .732$), and the right SF amygdala for both run 1, and combined runs ($\rho_{RSFR1}(13) = .263$, $p_{RSFR1} = .343$; $\rho_{RSFRC}(13) = .145$, $p_{RSFRC} = .606$).

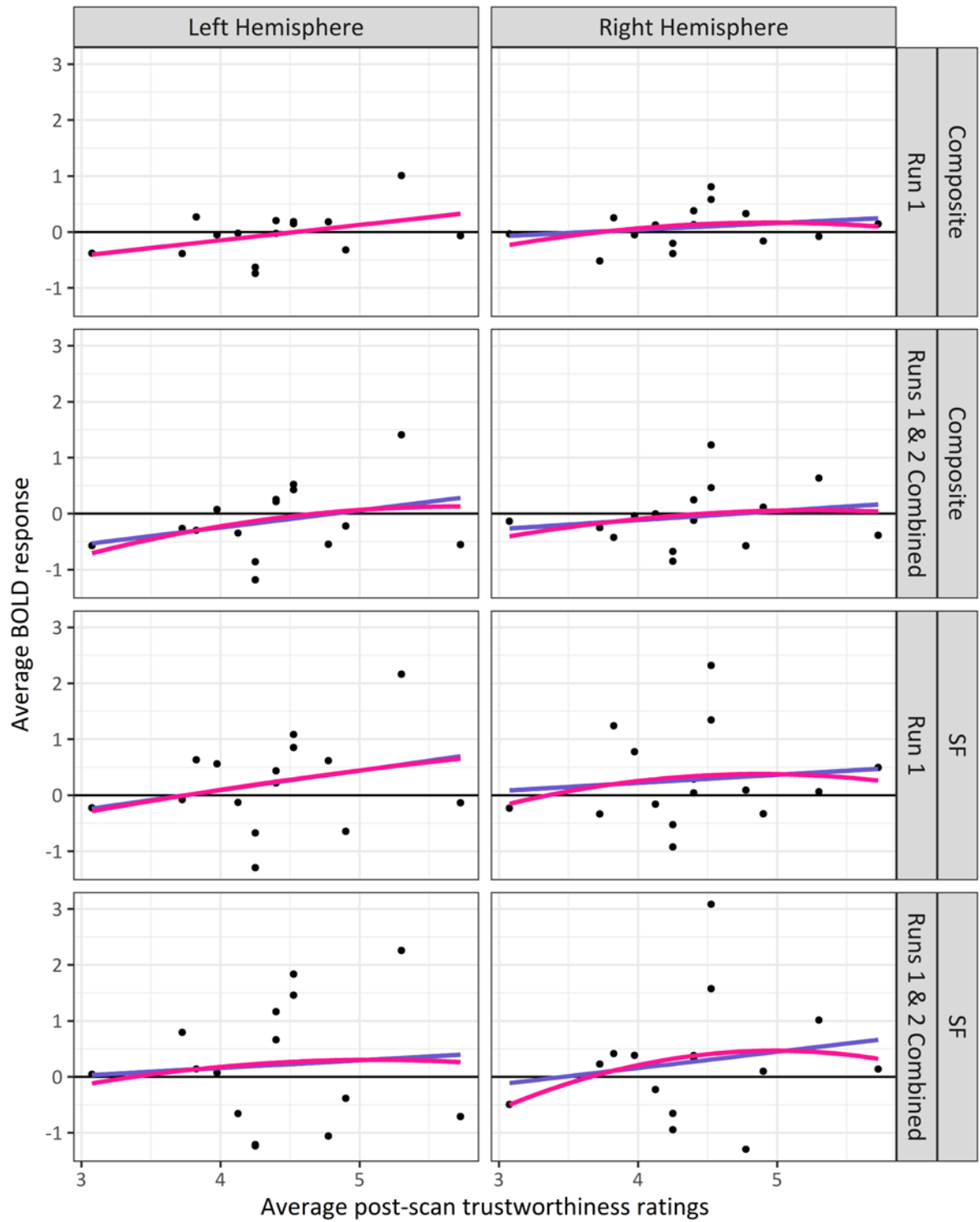


Figure 15: Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings of the male voices in the left composite amygdala for run 1 (row 1, left), the right composite amygdala for run 1 (row 1, right), the left composite amygdala for combined runs (row 2, left), the right composite amygdala for combined runs (row 2, right), the left SF amygdala for run 1 (row 3, left), the right SF amygdala for run 1 (row 3, right), the left SF amygdala for combined runs (row 4, left), and the right SF amygdala for combined runs (row 4, right). Linear (blue) and second-order polynomial (red) trendlines were added.

Additionally, a linear and a second level polynomial model were created bilaterally for each ROI in each investigated run (R1/ RC), and subsequently compared in a model comparison analysis. Both, the linear and the quadratic models are not a good fit for the averaged trustworthiness ratings of the male voices as only 0.62-17.36%, and 0.92-17.36% of the variance are explained by the linear and quadratic models respectively (Table 11).

Table 11: R^2 values for the linear model, the quadratic polynomial model, and a model comparison for male voice sex and post-scan trustworthiness ratings; separately for amygdala, hemisphere, and runs

Amygdala	Hemisphere	Run	R^2 linear model	F linear model (p-value)	R^2 quadratic model	F quadratic model (p-value)	F model comparison (p-value)
Composite	Left	1	0.1736	2.731 (.122)	0.1736	1.260 (.319)	<0.001 (.990)
		1&2	0.0960	1.380 (.261)	0.1086	0.731 (.502)	0.170 (.688)
	Right	1	0.0459	0.626 (.443)	0.0831	0.544 (.594)	0.486 (.499)
		1&2	0.0368	0.497 (.493)	0.0486	0.306 (.742)	0.148 (.707)
SF	Left	1	0.0736	1.034 (.328)	0.0742	0.481 (.630)	0.007 (.935)
		1&2	0.0062	0.082 (.779)	0.0092	0.056 (.946)	0.036 (.853)
	Right	1	0.0124	0.163 (.693)	0.0258	0.159 (.855)	0.164 (.692)
		1&2	0.0313	0.421 (.528)	0.0533	0.338 (.720)	0.278 (.608)

Linear and quadratic regression were then fitted to the individual responses in each ROI and a subsequent paired t-test, comparing the R^2 values for the quadratic polynomials and the linear model, showed the R^2 of the second-order polynomial to be significantly higher than the R^2 of the linear model for the left composite amygdala for both run 1 and combined runs ($t_{LCR1}(19) = -3.534$, $p_{LCR1} = .002$; $t_{LCRC}(19) = -3.800$, $p_{LCRC} = .001$), the right composite amygdala for both run 1 and combined runs ($t_{RCR1}(19) = -3.686$, $p_{RCR1} = .002$; $t_{RCRC}(19) = -2.116$, $p_{RCRC} = .048$), the left SF amygdala for both run 1 and combined runs ($t_{LSFR1}(19) = -3.181$, $p_{LSFR1} = .005$; $t_{LSFRC}(19) = -4.094$, $p_{LSFRC} < .001$), and the right SF amygdala for both run 1 and combined runs ($t_{RSFR1}(19) = -4.077$, $p_{RSFR1} < .001$; $t_{RSFRC}(19) = -2.885$, $p_{RSFRC} = .009$). For the individual participant's BOLD response in relation to the online validation data see Supplementary materials (Figure 30, and Figure 31).

Summary of results

The aim of Experiment 2b was to investigate the relationship between amygdala activation and perceptions of trustworthiness with an implicit PureTone detection task design low in attention and cognitive load. Behavioural results showed high correlations between the perceived trustworthiness ratings, gathered in the post-scan experiment, and the online validation ratings obtained in Experiment 1 for both male and female voices. This suggests that trustworthy voices were perceived as trustworthy, and untrustworthy voices as untrustworthy. Similarly to Experiment 2a, amygdala activation was then correlated to the trustworthiness ratings from the online validation experiment (Experiment 1), and the group averages of the post-scan trustworthiness ratings. No significant correlations between amygdala activation and perceived trustworthiness were found either for female or male voice stimuli. This suggests that whether perceived vocal trustworthiness can be traced in the amygdala depends on task design. Furthermore, a comparison on the individual data between linear and quadratic models showed that quadratic models were a better predictor of amygdala activation. This is in agreement with results from Experiment 2a, as well as findings from face research (i.e. Mattavelli et al., 2012). For further theoretical discussion of the results from Experiments 2a and 2b, see the next chapter “General discussion and conclusion” (page 69).

Comparison between tasks – fMRI data

To strengthen the comparison between tasks, the correlations from the 1-back and the PureTone detection task for all amygdala regions with significant Spearman's rank order correlation coefficients were compared.

When averaged group level amygdala activity was correlated to the online validation trustworthiness scores (Experiment 1), significant correlation values were only found for male voices. The regions included are the left and right superficial amygdala for run 1. Williams' T2 statistic (Steiger, 1980) was used to show that the two dependent correlations (1-back amygdala activity/ online validation trustworthiness scores; PureTone amygdala activity/ online validation trustworthiness scores) that shared a common variable (online validation trustworthiness scores) differed significantly (Table 12).

Table 12: Spearman's rank order correlation coefficients for the group level amygdala activity and the online validation trustworthiness scores (Experiment 1) in the 1-back and the PureTone task, and William's T2 statistic comparing the two tasks.

Amygdala region & run	Spearman's rho (1-back task)	p-value (1-back task)	Spearman's rho (PureTone task)	p-value (PureTone task)	Williams' T2	p-value for Williams' T2
Left superficial amygdala run 1	-.525	.044	.361	.187	3.33	.006
Right superficial amygdala run 1	-.525	.044	.318	.248	3.22	.007

Correlating the averaged group level amygdala activity and the post-scan trustworthiness responses, significant Spearman's rank order correlations were found only in male voices for the left composite amygdala run 1, the left composite amygdala for combined runs, the left superficial amygdala run 1, and the right superficial amygdala run 1. A Fisher's r-z transformation statistic (Steiger, 1980) was used to show that the two independent correlations (1-back amygdala activity/ 1-back post-scan trustworthiness responses; PureTone amygdala activity/ PureTone post-scan trustworthiness responses) were significantly different from one another (Table 13).

Table 13: Spearman's rank order correlation coefficients for the group level amygdala activity and the post-scan behavioural responses in the 1-back and the PureTone task, and Fisher's r-z transformation comparing the two tasks.

Amygdala region & run	Spearman's rho (1-back task)	p-value (1-back task)	Spearman's rho (PureTone task)	p-value (PureTone task)	Fisher's z	p-value for Fisher's z
Left composite amygdala run 1	-.604	.017	.326	.236	2.54	.011
Left composite amygdala runs 1 and 2 combined	-.517	.049	.342	.212	2.27	.023
Left superficial amygdala run 1	-.624	.013	.242	.388	2.40	.017
Right superficial amygdala run 1	-.622	.013	.263	.343	2.44	.015

General discussion and conclusion

Across two experiments, we investigated the relationship between amygdala activation and perceived trustworthiness. In Experiment 1, we pre-validated male and female voices on perceived trustworthiness, and selected 15 stimuli for each voice sex to be used in the fMRI experiments (Experiment 2a and 2b). We hypothesised that amygdala activation and perceived trustworthiness would correlate negatively regardless of voice sex, and independent of task (i.e. 1-back task, PureTone detection task). The hypotheses were only partially confirmed, as significant correlation values were found for male but not female voices in the 1-back experiment (Experiment 2a), and no other significant relationships between amygdala activation and perceived trustworthiness in either voice sex was established for the PureTone experiment (Experiment 2b). Results held true for behavioural consensus data obtained from the online validation experiment (Experiment 1) as well as the data gathered from the fMRI participants in the post-scan behavioural experiment. A secondary aim was to explore the nature of the relationship between amygdala activation and perceived trustworthiness further. It was established that quadratic models predicted amygdala activation significantly better than linear models despite not explaining much variance of the data.

The result of a significant negative relationship between amygdala activation and perceived trustworthiness was obtained for male but not female voices. Non-findings are in agreement with Hensel et al. (2015) who did not list the amygdala as a region of interest involved in explicit social judgements. The significant findings for male voices mirror results of amygdala activation in relation to perceived facial trustworthiness (Engell et al., 2007; Freeman et al., 2014; Mattavelli et al., 2012; Said et al., 2009; Todorov, Baron, et al., 2008; Todorov & Engell, 2008; Winston et al., 2002). To be more precise, activation has been located in the SF subdivision of the amygdala which is in line with a meta-analysis in faces (Bzdok et al., 2011). It is slightly puzzling this study produced significant results for male but not for female voices. This cannot be due to voice sex differences in perceived trustworthiness as the ranges of male and female voices did not differ significantly (Experiment 1) unless effect sizes in female voices are more subtle, and would require a greater range of trustworthiness ratings. Neither can a reason be sought in the differences of behavioural perceptions of the fMRI participants, as their perceived trustworthiness ratings, gathered in the post-scan

experiment, correlated highly with those obtained in the online validation experiment (Experiment 1). One theoretical explanation can be found in the approach/ avoidance theorem suggested by Evolutionary Theory and/ or by Oosterhof and Todorov's overgeneralisation model (2008). Since there is a close connection between the concepts of trustworthiness and threat (Oosterhof & Todorov, 2008), and evolutionary findings show that males are generally perceived as more threatening than females (Puts, 2010, 2016; Puts et al., 2012), it is possible that perceived trustworthiness can be traced in the amygdala for male but not for female voices. It needs to be emphasized that Hensel et al. (2015) did not analyse data separately for each voice sex. If differences between male and female voice sex truly exist, combining both voice sexes reduces the chance of detecting neural areas involved in the perception of trustworthiness.

It is also possible that the amygdala is a "relevance detector" (Pernet et al., 2015; Sander et al., 2003) rather than being involved in processing emotion or social information.

Trustworthiness might be less salient for female than for male voices which could explain the absence of significant correlations for the female voice sex. Similarly, habituation or adaptation effects (Belin & Zatorre, 2003; Leopold, O'Toole, Vetter, & Blanz, 2001; Rankin et al., 2009) might be sought as a further explanation for the absence of significant results not only for female in relation to male voices but also comparing results from run 1 and combined runs 1 and 2. Amygdala activity has been shown to decrease rapidly to repeatedly presented stimuli (Breiter et al., 1996; Fruehholz & Grandjean, 2013; Plichta et al., 2014; Wiethoff, Wildgruber, Grodd, & Ethofer, 2009). Despite randomising the order of the 15 voice stimuli per block, habituation or adaption effects might still have occurred. A future study should investigate this by analysing single blocks (or the first few blocks) within run 1, however the number of stimuli within each block should be increased.

Contrary to our hypothesis, correlations between amygdala activation and perceived trustworthiness were task-dependent as significant values were reported for the 1-back but not the PureTone detection task. This could be due to reasons of attention and cognitive load. The 1-back task required a higher degree of attention and cognitive load to be able to compare the present with the previous stimulus in order to decide whether they were different or not whereas attention and cognitive load in the PureTone detection task were lowered by requesting participants to simply decide whether the stimulus was a voice or not.

The PureTone detection task was created based on the red spot detection task used by Mattavelli et al. (2012), however the authors reported a relationship between amygdala activation and perceived trustworthiness whereas no significant correlations can be reported in the current study. An explanation of using unnatural attention keepers does not apply as both the red dot and the beep tone are artificially different, and the faces and voices should have been perceived by participants even if so on a subconscious level. A potential reason as to why the PureTone task did not elicit amygdala activation could be sought however in task difficulty, as under-stimulation can manifest in failures in attentional allocation (Raffaelli, Mills, & Christoff, 2017).

Furthermore, Mattavelli et al. (2012) and the current study differed in stimuli selection and task design. By computer manipulation, the authors created a matrix of faces changing alongside dimensions of trustworthiness and gender, allowing them to select 10 faces in each block varying on a gender but not a trustworthiness dimension. Their experiment comprised of four blocks of different trustworthiness levels (a total of 40 different faces) with each block being repeated five times. In the current study, for each voice sex, we employed a total of 15 voices within each block which were evenly distributed alongside the trustworthiness dimension. Each block was then repeated 10 times in each of the two runs. A problem for the current study could have been the extensive repetitions of the vocal stimuli, given that the amygdala has been shown to elicit responses to novel stimuli, but more importantly, that these habituate rapidly if the stimulus is repeated (Kosaka et al., 2003; LeDoux, 2007; Sander et al., 2003; Schwarz & Hassebrauck, 2012). A solution for future research could lie in a different way of stimuli selection (i.e. 10 voices in each category of high, medium, and low trustworthiness), or in morphing techniques (Bestelmeyer et al., 2012; Bruckert et al., 2010; Watson et al., 2013). This would maintain more control over trustworthiness levels, as well as creating multiple voice stimuli for each trustworthiness category to enable reducing repetitions of the same voice. If task design is seemingly so imperative for detecting amygdala activation, and given that a clear distinction in neural activation pattern between explicit and implicit tasks in face research has yet to be established, that would provide an additional explanation as to why Hensel et al. (2015) did not report the amygdala a key region involved in perceived trustworthiness. Also, since Mattavelli et al. (2012) created stimuli along four distinct trustworthiness categories, maybe

selecting vocal stimuli with approximately equal distances to one another could have resulted in differences being too minimal to be detected. However, these last two notions remain speculative.

As a secondary aim was to identify an activation pattern, linear and quadratic model approaches were compared. Across both experimental tasks, the current study found non-linear/ quadratic effects a better predictor of amygdala activation than linear models. This is in agreement with Mattavelli et al. (2012), and partially with Freeman et al. (2014) who detected linear and quadratic responses depending on different subdivisions of the amygdala. In face research, negative linear response patterns in the amygdala have been understood in relation to valence and threat detection (Engell et al., 2007; LeDoux, 2003, 2007, 2012; Todorov & Engell, 2008; Todorov, Said, et al., 2008), whereas non-linear patterns have been frequently interpreted as relevant for salience or motivation (Adolphs, 2010; Freeman et al., 2014; Todorov et al., 2013). With finding quadratic effects in this study, a similar interpretation to the ones already used in face research is attempted. Assuming a concept of self-preservation, a person with an untrustworthy voice wants to be avoided (the amygdala could be seen as a warning system) whereas a highly trustworthy-sounding person needs to be approached for social-motivational reasons. Another, not mutually exclusive, interpretation to the one addressed above can be seen in Todorov's typicality framework (Said et al., 2010; Todorov, 2012). The authors stated that the amygdala responds more strongly the less typical the face stimuli were. Given the current study's findings, it can be speculated that this holds true for voice stimuli as well, however to confirm this interpretation the voice stimuli used here should be re-validated for vocal typicality. This would also allow to assess whether the concept of a vocal prototype in relation to social judgements exists (see also Latinus & Belin, 2011).

Whilst we found quadratic models to be significantly better predictors of amygdala activation, the explained variance was low. This is in sharp contrast to Mattavelli et al. (2012) who reported R^2 values of 0.63 for their quadratic polynomial models. It could be speculated that this is either due to Mattavelli et al. (2012) combining the data for the bilateral hemispheres whereas in this study data from the left and right hemisphere were analysed separately, or that the models in the current study were simply underspecified. Adding further predictors, such as listener sex, might aid to improve the models, however this was

not explored further in this study given the already small sample size, and that behavioural personality perceptions do not differ between listener sex (Bruckert et al., 2010; Mahrholz et al., under review). Another reason as to why quadratic polynomials explain more of the data variance in face studies than vocal models could be that we live in a predominantly visual world. Multi-modal behavioural research has shown that perceived trustworthiness is driven either by the face (Mileva et al., 2017) or the integration of facial and vocal modalities (Rezlescu et al., 2015). If this holds true for amygdala activation in relation to perceived trustworthiness as well, a multimodal approach could lead to models that explain more variance than a single facial or vocal model could contribute on its own.

Further to some of the limitations already conferred, one caveat to our study is a low sample size for the experiments 2a and 2b, which is typical for neuroscience studies. Low sample size frequently denotes low power of a study, which does not allow for strong conclusions. Both experiment 2a and 2b had each 20 participants recruited, but pooling of the data sets to increase power was prevented given that results were task-dependent as discussed before. Low statistical power suggests that the chance of discovering true effects is small, the probability that an observed effect is a true effect is low, and if the effect is indeed a true effect, the magnitude of this effect might be exaggerated (Button et al., 2013). This is also an issue for the studies by Mattavelli et al. (2012) and Freeman et al. (2014), whose results were based on 20, and 16 participants respectively. This suggests that the high R^2 value reported by Mattavelli et al. (2012) could be exaggerated. With this caveat in mind, both fMRI experiments in this study had been designed with two runs of each voice sex to increase power, however as described previously that could have prevented finding effects due to extensive repetitions of the same voices.

Due to our study design, exploratory rather than confirmatory in nature, the significant findings in male voices required Bonferroni corrections for multiple comparisons which rendered all effects non-significant. To investigate whether the found effects are true effects, the 1-back experiment should be replicated on an independent data set with more participants (Nosek et al., 2015). However, future research should also attempt exploratory multivariate whole brain analyses as this would assist in establishing additional neural areas involved in the processing of vocal trustworthiness.

In summary, it is proposed that neural activation within the amygdala, more precisely in the SF subdivision, is negatively correlated to vocal trustworthiness, reflecting a similar relationship found with faces. The relationship is stronger for the male than the female voice sex suggesting more subtle effects for female voices. Results are also shown to be task-dependent. Multivariate whole brain analyses would aid in establishing additional neural areas involved in processing vocal trustworthiness. Overall, the amygdala is sensitive to modulations in socially relevant vocal characteristics related to approach/avoidance decisions.

References

- Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Year in Cognitive Neuroscience 2010*, 1191, 42-61. doi:10.1111/j.1749-6632.2010.05445.x
- Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature*, 393(6684), 470-474. doi:10.1038/30982
- Allport, G. W., & Cantril, H. (1934). Judging personality from voice. *The Journal of Social Psychology*, 5, 37-55. doi:10.1080/00224545.1934.9921582
- Aronovitch, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of social psychology*, 99(2), 207-220.
- Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, 6(2), 269-278. doi:10.1037/1528-3542.6.2.269
- Baugh, J. (2000). Racial identifications by speech. *American Speech*, 75(4), 362-364. doi:10.1215/00031283-75-4-362
- Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport*, 14(16), 2105-2109. doi:10.1097/00001756-200311140-00019
- Bestelmeyer, P. E. G., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F., & Belin, P. (2012). Implicitly Perceived Vocal Attractiveness Modulates Prefrontal Cortex Activity. *Cerebral Cortex*, 22(6), 1263-1270. doi:10.1093/cercor/bhr204
- Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, 82(1), 55-59. doi:10.1016/j.anbehav.2011.03.024
- Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., . . . Rosen, B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, 17(5), 875-887. doi:10.1016/s0896-6273(00)80219-6
- Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., . . . Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology*, 20(2), 116-120.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365-376. doi:10.1038/nrn3475
- Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., . . . Eickhoff, S. B. (2011). ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure & Function*, 215(3-4), 209-223. doi:10.1007/s00429-010-0287-4

- Cabeza, R., Anderson, N. D., Kester, J., & McIntosh, A. R. (2002). Hemispheric Asymmetry Reduction in Old Adults (HAROLD): Evidence for the compensation hypothesis. *Journal of Cognitive Neuroscience*, 125-125.
- Castle, E., Eisenberger, N. I., Seeman, T. E., Moons, W. G., Boggero, I. A., Grinblatt, M. S., & Taylor, S. E. (2012). Neural and behavioral bases of age differences in perceptions of trust. *Proceedings of the National Academy of Sciences*, 109(51), 20848-20852. doi:10.1073/pnas.1218518109
- Davis, S. W., Dennis, N. A., Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2008). Que PASA? The posterior-anterior shift in aging. *Cerebral Cortex*, 18(5), 1201-1209. doi:10.1093/cercor/bhm155
- DeBruine, L., Jones, B. C., Frederick, D. A., Haselton, M. G., Penton-Voak, I. S., & Perrett, D. I. (2010). Evidence for Menstrual Cycle Shifts in Women's Preferences for Masculinity: A Response to Harris (in press) "Menstrual Cycle and Facial Preferences Reconsidered". *Evolutionary Psychology*, 8(4), 768-775.
- Dixon, J. A., Mahoney, B., & Cocks, R. (2002). Accents of guilt? Effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology*, 21(2), 162-168. doi:10.1177/02627x02021002004
- Dunbar, N. E., & Burgoon, J. K. (2005). Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2), 207-233.
- Efran, M. G. (1974). The effect of physical appearance on the judgment of guilt, interpersonal attraction, and severity of recommended punishment in a simulated jury task. *Journal of Research in Personality*, 8(1), 45-54. doi:10.1016/0092-6566(74)90044-0
- Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, 19(9), 1508-1519.
- Freeman, J. B., Stoller, R. M., Ingbretsen, Z. A., & Hehman, E. A. (2014). Amygdala Responsivity to High-Level Social Information from Unseen Faces. *Journal of Neuroscience*, 34(32), 10573-10581. doi:10.1523/jneurosci.5063-13.2014
- Fruehholz, S., & Grandjean, D. (2013). Amygdala subregions differentially respond and rapidly adapt to threatening voices. *Cortex*, 49(5), 1394-1403. doi:10.1016/j.cortex.2012.08.003
- Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, 21(3), 177-182. doi:10.1177/0963721412445309
- Gorn, G. J., Jiang, Y., & Johar, G. V. (2008). Babyfaces, trait inferences, and company evaluations in a public relations crisis. *Journal of Consumer Research*, 35(1), 36-49. doi:10.1086/529533

- Grady, C. (2012). The cognitive neuroscience of ageing. *Nature Reviews Neuroscience*, 13(7), 491-505. doi:10.1038/nrn3256
- Gutchess, A. (2014). Plasticity of the aging brain: New directions in cognitive neuroscience. *Science*, 346(6209), 579-582. doi:10.1126/science.1254604
- Hensel, L., Bzdok, D., Mueller, V. I., Zilles, K., & Eickhoff, S. B. (2015). Neural Correlates of Explicit Social Judgments on Vocal Stimuli. *Cerebral Cortex*, 25(5), 1152-1162. doi:10.1093/cercor/bht307
- Herzog, H. (1933). Stimme und Persönlichkeit (Voice and Personality). *Zeitschrift Fur Psychologie Und Physiologie Der Sinnesorgane*, 130(3-5), 300-369.
- Hughes, S. M., & Rhodes, B. C. (2010). Making age assessments based on voice: The impact of the reproductive viability of the speaker. *Journal of Social, Evolutionary, and Cultural Psychology*, 4(4), 290-304. doi:10.1037/h0099282
- Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, 121(3), 313-323. doi:10.1016/j.cognition.2011.08.001
- Jiang, X., & Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. *Cortex*(0). doi:<http://dx.doi.org/10.1016/j.cortex.2015.02.002>
- Jones, B. C., DeBruine, L. M., Perrett, D. I., Little, A. C., Feinberg, D. R., & Smith, M. J. L. (2008). Effects of menstrual cycle phase on face preferences. *Archives of Sexual Behavior*, 37(1), 78-84. doi:10.1007/s10508-007-9268-y
- Kennedy, D. P., Gläscher, J., Tyszka, J. M., & Adolphs, R. (2009). Personal space regulation by the human amygdala. *Nature Neuroscience*, 12(10), 1226-1227. doi:10.1038/nn.2381
- Klofstad, C. A., Anderson, R., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PloS one*, 10(8), e0133779. doi:10.1371/journal.pone.0133779
- Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings. Biological Sciences / The Royal Society*, 279(1738), 2698-2704. doi:10.1098/rspb.2012.0311
- Kosaka, H., Omori, M., Iidaka, T., Murata, T., Shimoyama, T., Okada, T., . . . Wada, Y. (2003). Neural substrates participating in acquisition of facial familiarity: an fMRI study. *Neuroimage*, 20(3), 1734-1742. doi:10.1016/s1053-8119(03)00447-6
- Koscik, T. R., & Tranel, D. (2011). The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia*, 49(4), 602-611. doi:10.1016/j.neuropsychologia.2010.09.023

- Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, 38(6), 618-625. doi:10.1016/S0022-1031(02)00510-3
- Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, 21(4), R143-R145.
- Latinus, M., & Belin, P. (2012). Perceptual Auditory Aftereffects on Voice Identity Using Brief Vowel Stimuli. *Plos One*, 7(7). doi:10.1371/journal.pone.0041384
- LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and molecular neurobiology*, 23(4-5), 727-738.
- LeDoux, J. (2007). The amygdala. *Current Biology*, 17(20), R868-R874. doi:10.1016/j.cub.2007.08.005
- LeDoux, J. (2012). Rethinking the Emotional Brain. *Neuron*, 73(4), 653-676. doi:10.1016/j.neuron.2012.02.004
- Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level after effects. *Nature Neuroscience*, 4(1), 89-94. doi:10.1038/82947
- Mahrholz, G., Belin, P., & McAleer, P. (under review). First impressions of a speaker's personality show stability across short temporal durations and are unaffected by content. *PLOS ONE*.
- Mattavelli, G., Andrews, T. J., Asghar, A. U., Towler, J. R., & Young, A. W. (2012). Response of face-selective brain regions to trustworthiness and gender of faces. *Neuropsychologia*, 50(9), 2205-2211.
- McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS ONE*, 9(3), e90779. doi:10.1371/journal.pone.0090779
- McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience*, 27, 1-28. doi:10.1146/annurev.neuro.27.070203.144157
- Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience*, 8(3), 285-299. doi:10.1093/scan/nsr090
- Mileva, M., Tompkinson, J. A., Watt, D., & Burton, A. M. (2017). Audiovisual Integration in Social Evaluation. *Journal of Experimental Psychology: Human Perception and Performance*.
- Moyse, E., Beaufort, A., & Brédart, S. (2014). Evidence for an own-age bias in age estimation from voices in older persons. *European Journal of Ageing*, 11(3), 241-247. doi:10.1007/s10433-014-0305-0

- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422-1425. doi:10.1126/science.aab2374
- Nyberg, L., Lovden, M., Riklund, K., Lindenberger, U., & Backman, L. (2012). Memory aging and brain maintenance. *Trends in Cognitive Sciences*, 16(5), 292-305. doi:10.1016/j.tics.2012.04.005
- Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, 34(2), 83-110. doi:10.1007/s10919-009-0082-1
- Olivola, C. Y., & Todorov, A. (2010b). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, 46(2), 315-324. doi:10.1016/j.jesp.2009.12.002
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 105(32), 11087-11092. doi:10.1073/pnas.0805664105
- Pear, T. H. (1931). *Voice and Personality*. London: Chapman & Hall.
- Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., . . . Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, 119, 164-174. doi:10.1016/j.neuroimage.2015.06.050
- Perrachione, T. K., & Ghosh, S. S. (2013). Optimized design and analysis of sparse-sampling fMRI experiments. *Frontiers in Neuroscience*, 7. doi:10.3389/fnins.2013.00055
- Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., . . . Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*, 95, 89-99. doi:10.1016/j.anbehav.2014.06.011
- Pisanski, K., Jones, B. C., Fink, B., O'Connor, J. J. M., DeBruine, L. M., Röder, S., & Feinberg, D. R. (2016). Voice parameters predict sex-specific body morphology in men and women. *Animal Behaviour*, 112, 13-22. doi:10.1016/j.anbehav.2015.11.008
- Plichta, M. M., Grimm, O., Morgen, K., Mier, D., Sauer, C., Haddad, L., . . . Meyer-Lindenberg, A. (2014). Amygdala habituation: A reliable fMRI phenotype. *Neuroimage*, 103, 383-390. doi:10.1016/j.neuroimage.2014.09.059
- Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, 18(1), 10-30. doi:10.1177/0261927x99018001002
- Puts, D. A. (2010). Beauty and the beast: mechanisms of sexual selection in humans. *Evolution and Human Behavior*, 31(3), 157-175. doi:10.1016/j.evolhumbehav.2010.02.005

- Puts, D. A. (2016). Human sexual selection. *Current Opinion in Psychology*, 7, 28-32. doi:10.1016/j.copsyc.2015.07.011
- Puts, D. A., Apicella, C. L., & Cárdenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings. Biological Sciences / The Royal Society*, 279(1728), 601-609. doi:10.1098/rspb.2011.0829
- Quadflieg, S., Todorov, A., Laguesse, R., & Rossion, B. (2012). Normal face-based judgements of social characteristics despite severely impaired holistic face processing. *Visual Cognition*, 20(8), 865-882. doi:10.1080/13506285.2012.707155
- Raffaelli, Q., Mills, C., & Christoff, K. (2017). The knowns and unknowns of boredom: a review of the literature. *Experimental brain research*. doi:10.1007/s00221-017-4922-7
- Rakic, T., Steffens, M. C., & Mummendey, A. (2011). When it matters how you pronounce it: The influence of regional accents on job interview outcome. *British Journal of Psychology*, 102, 868-883. doi:10.1111/j.2044-8295.2011.02051.x
- Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., . . . Thompson, R. F. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory*, 92(2), 135-138. doi:10.1016/j.nlm.2008.09.012
- Reuter-Lorenz, P. A., & Cappell, K. A. (2008). Neurocognitive aging and the compensation hypothesis. *Current Directions in Psychological Science*, 17(3), 177-182. doi:10.1111/j.1467-8721.2008.00570.x
- Reuter-Lorenz, P. A., & Park, D. C. (2010). Human Neuroscience and the Aging Mind: at Old Problems A New Look. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences*, 65(4), 405-415. doi:10.1093/geronb/gbq035
- Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant Voices and Attractive Faces: The Contribution of Visual and Auditory Information to Integrated Person Impressions. *Journal of Nonverbal Behavior*, 39(4), 355-370. doi:10.1007/s10919-015-0214-8
- Roozendaal, B., McEwen, B. S., & Chattarji, S. (2009). Stress, memory and the amygdala. *Nature Reviews Neuroscience*, 10(6), 423-433. doi:10.1038/nrn2651
- Said, C. P., Baron, S. G., & Todorov, A. (2009). Nonlinear Amygdala Response to Face Trustworthiness: Contributions of High and Low Spatial Frequency Information. *Journal of Cognitive Neuroscience*, 21(3), 519-528. doi:10.1162/jocn.2009.21041
- Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia*, 48(12), 3596-3605. doi:10.1016/j.neuropsychologia.2010.08.009

- Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society*, 16(5), 754-760. doi:10.1017/s1355617710000706
- Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the Neurosciences*, 14(4), 303-316.
- Scherer, K. R. (1972). Judging personality from voice: a cross-cultural approach to an old issue in interpersonal perception. *Journal of Personality*, 40(2), 191-&. doi:10.1111/j.1467-6494.1972.tb00998.x
- Schroeder, J., & Epley, N. (2015). The Sound of Intellect: Speech Reveals a Thoughtful Mind, Increasing a Job Candidate's Appeal. *Psychological Science*, 26(6), 877-891. doi:10.1177/0956797615572906
- Schwarz, S., & Hassebrauck, M. (2012). Sex and age differences in mate-selection preferences. *Human Nature*, 23(4), 447-466. doi:10.1007/s12110-012-9152-x
- St Jacques, P. L., Dolcos, F., & Cabeza, R. (2009). Effects of Aging on Functional Connectivity of the Amygdala for Subsequent Memory of Negative Pictures: A Network Analysis of Functional Magnetic Resonance Imaging Data. *Psychological Science*, 20(1), 74-84. doi:10.1111/j.1467-9280.2008.02258.x
- Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2), 245-251. doi:10.1037/0033-2909.87.2.245
- Stewart, G. L., Dustin, S. L., Barrick, M. R., & Darnold, T. C. (2008). Exploring the handshake in employment interviews. *Journal of Applied Psychology*, 93(5), 1139-1146. doi:10.1037/0021-9010.93.5.1139
- Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: ambient images generate a three-dimensional model. *Cognition*, 127(1), 105-118. doi:10.1016/j.cognition.2012.12.001
- Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, 33(3), 210-216. doi:10.1016/j.evolhumbehav.2011.09.004
- Todorov, A. (2012). The role of the amygdala in face perception and evaluation. *Motivation and Emotion*, 36(1), 16-26. doi:10.1007/s11031-011-9238-5
- Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience*, 3(2), 119-127. doi:10.1093/scan/nsn009
- Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, 25(3), 395-410. doi:10.1080/02643290802044996

- Todorov, A., & Engell, A. D. (2008). The role of the amygdala in implicit evaluation of emotionally neutral faces. *Social Cognitive and Affective Neuroscience*, 3(4), 303-312. doi:10.1093/scan/nsn033
- Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308(5728), 1623-1626. doi:10.1126/science.1110589
- Todorov, A., Mende-Siedlecki, P., & Dotsch, R. (2013). Social judgments from faces. *Current Opinion in Neurobiology*, 23(3), 373-380. doi:10.1016/j.conb.2012.12.010
- Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, 27(6), 813-833. doi:10.1521/soco.2009.27.6.813
- Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, 12(12), 455-460. doi:10.1016/j.tics.2008.10.001
- Todorov, A., Said, C. P., Oosterhof, N. N., & Engell, A. D. (2011). Task-invariant Brain Responses to the Social Value of Faces. *Journal of Cognitive Neuroscience*, 23(10), 2766-2781. doi:10.1162/jocn.2011.21616
- Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 111(32), E3353-E3361. doi:10.1073/pnas.1409860111
- Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12), 585-594. doi:10.1016/j.tics.2005.10.011
- Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2013). Dissociating task difficulty from incongruence in face- voice emotion integration. *Frontiers in Human Neuroscience*, 7. doi:10.3389/fnhum.2013.00744
- Wiethoff, S., Wildgruber, D., Grodd, W., & Ethofer, T. (2009). Response and habituation of the amygdala during processing of emotional prosody. *Neuroreport*, 20(15), 1356-1360. doi:10.1097/WNR.0b013e328330eb83
- Willis, J., & Todorov, A. (2006). First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science (Wiley-Blackwell)*, 17(7), 592-598. doi:10.1111/j.1467-9280.2006.01750.x
- Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, 5(3), 277-283. doi:10.1038/nn816
- Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, 17(6), 263-271. doi:10.1016/j.tics.2013.04.004

- Zalla, T., & Sperduti, M. (2013). The amygdala and the relevance detection theory of autism: an evolutionary perspective. *Frontiers in Human Neuroscience*, 7. doi:10.3389/fnhum.2013.00894
- Zebrowitz, L. A. (1996). Physical appearance as a basis of stereotyping. *Stereotypes and stereotyping*, 79-120.
- Zebrowitz, L. A., & Collins, M. A. (1997). Accurate Social Perception at Zero Acquaintance: The Affordances of a Gibsonian Approach. *Personality & Social Psychology Review (Lawrence Erlbaum Associates)*, 1(3), 204.
- Zebrowitz, L. A., & Montepare, J. M. (2008). Social Psychological Face Perception: Why Appearance Matters. *Social And Personality Psychology Compass*, 2(3), 1497-1497.
- Zuckerman, M., & Driver, R. E. (1989). What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of Nonverbal Behavior*, 13(2), 67-82. doi:10.1007/BF00990791

Supplementary materials

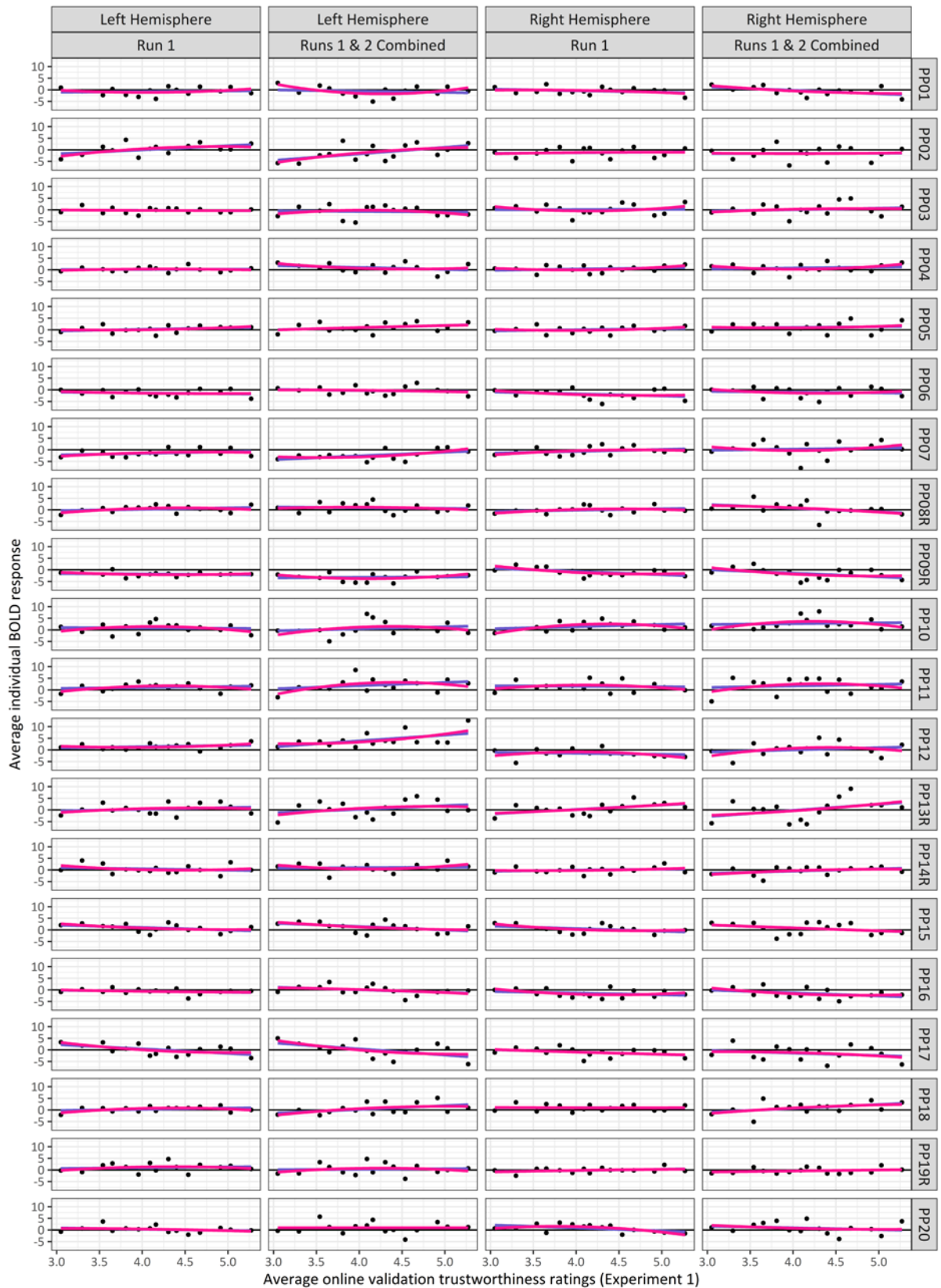


Figure 16: Female voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

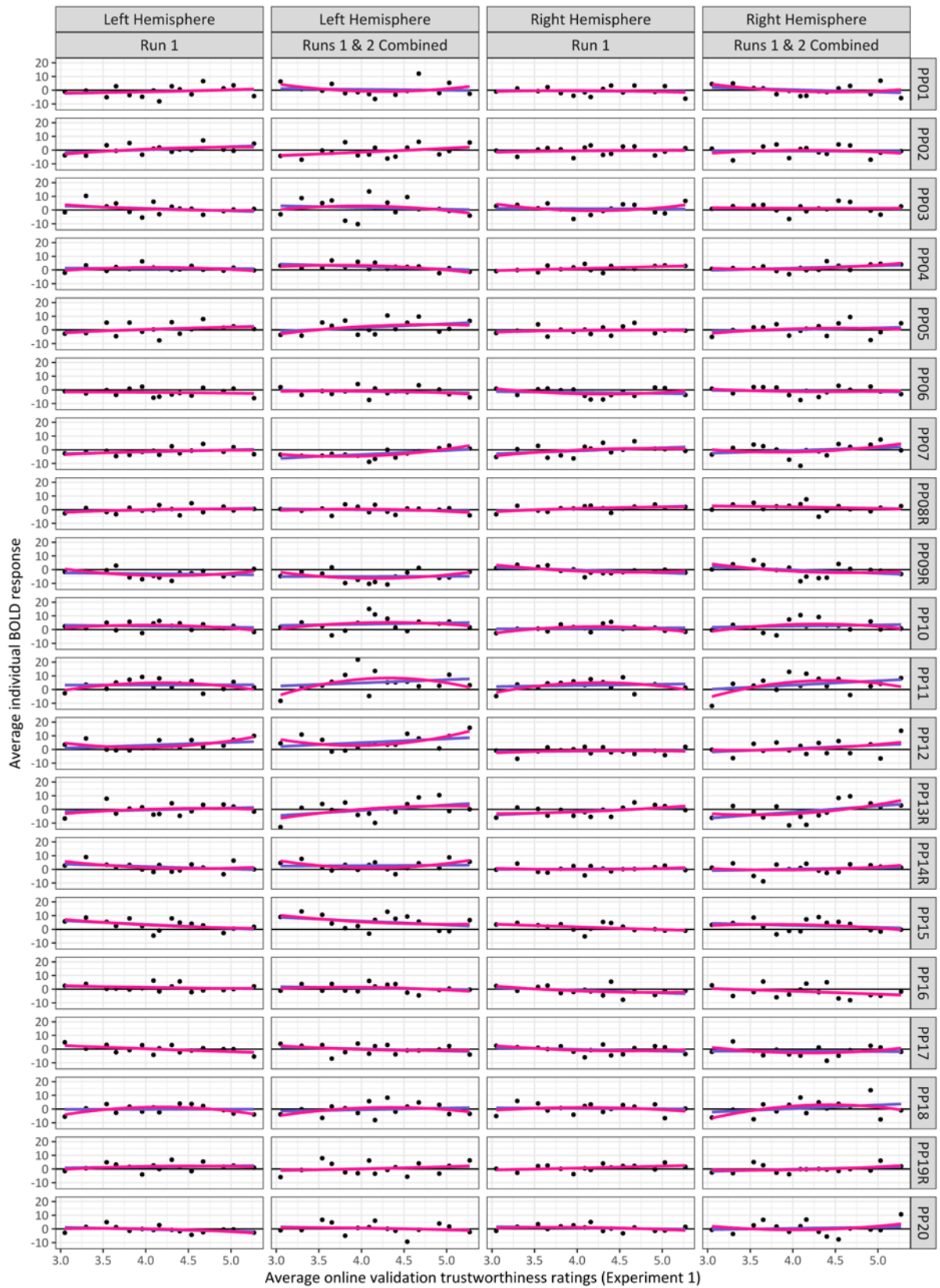


Figure 17: Female voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

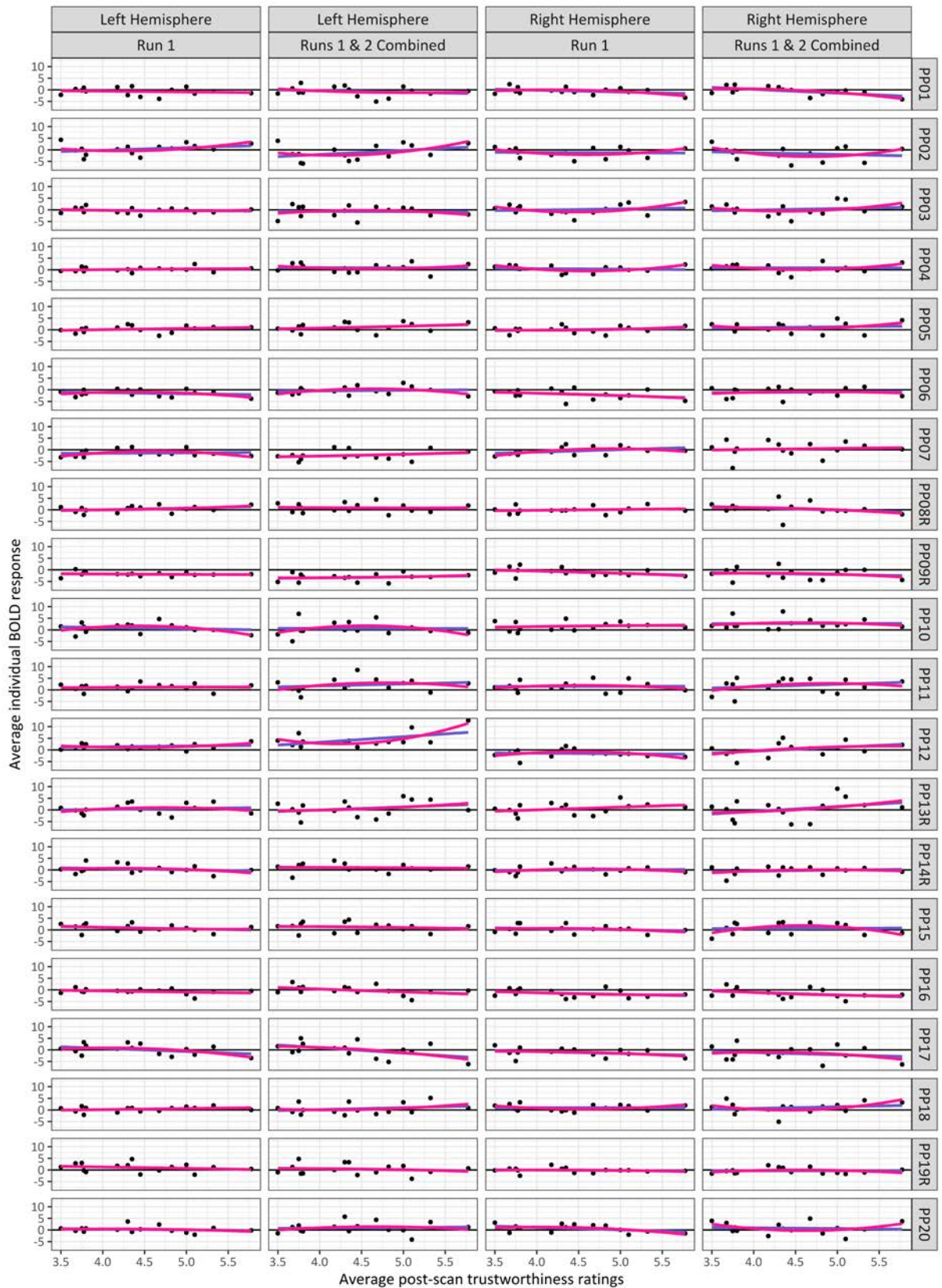


Figure 18: Female voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

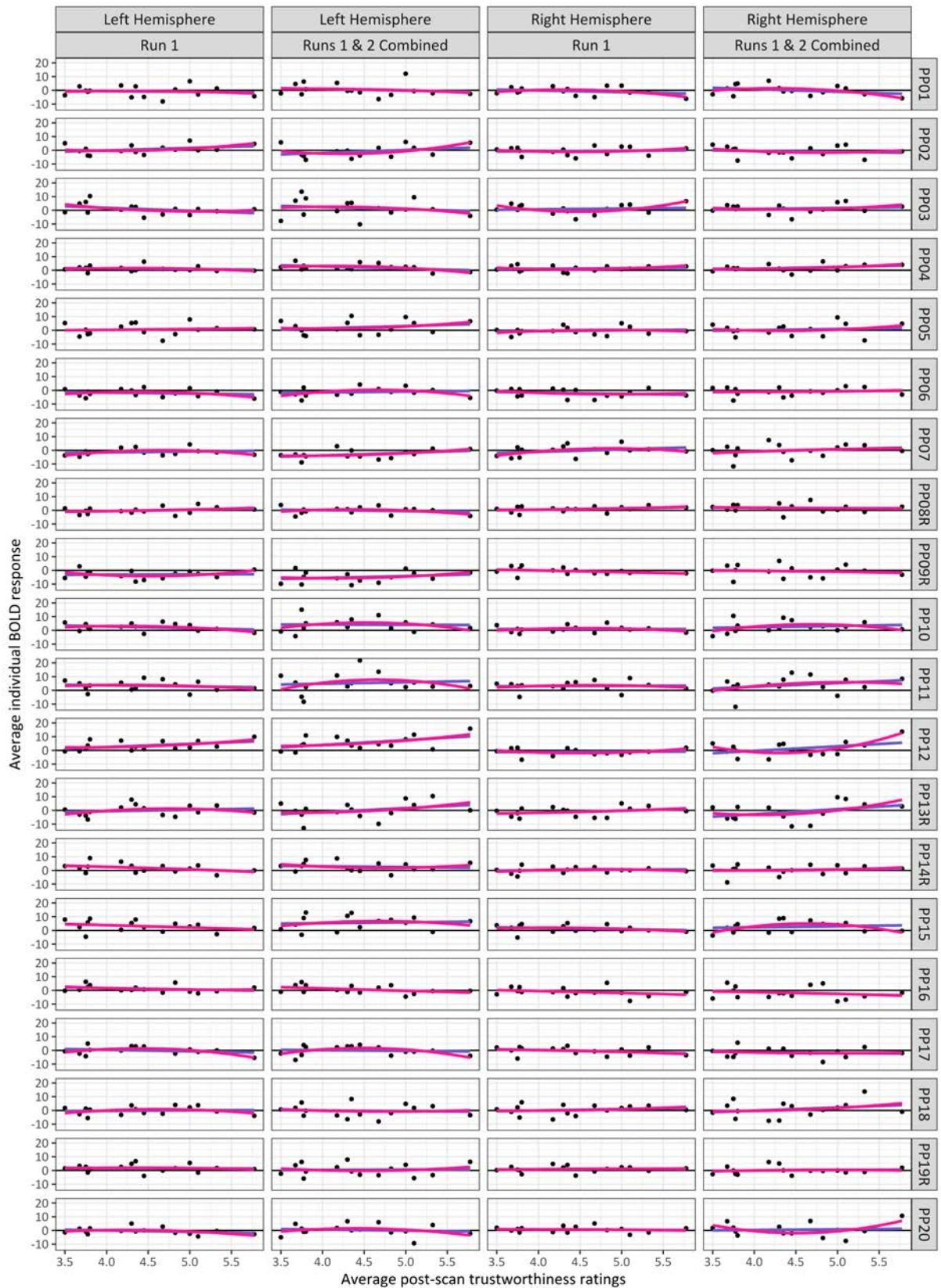


Figure 19: Female voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order Polynomial (red) trendlines were added.

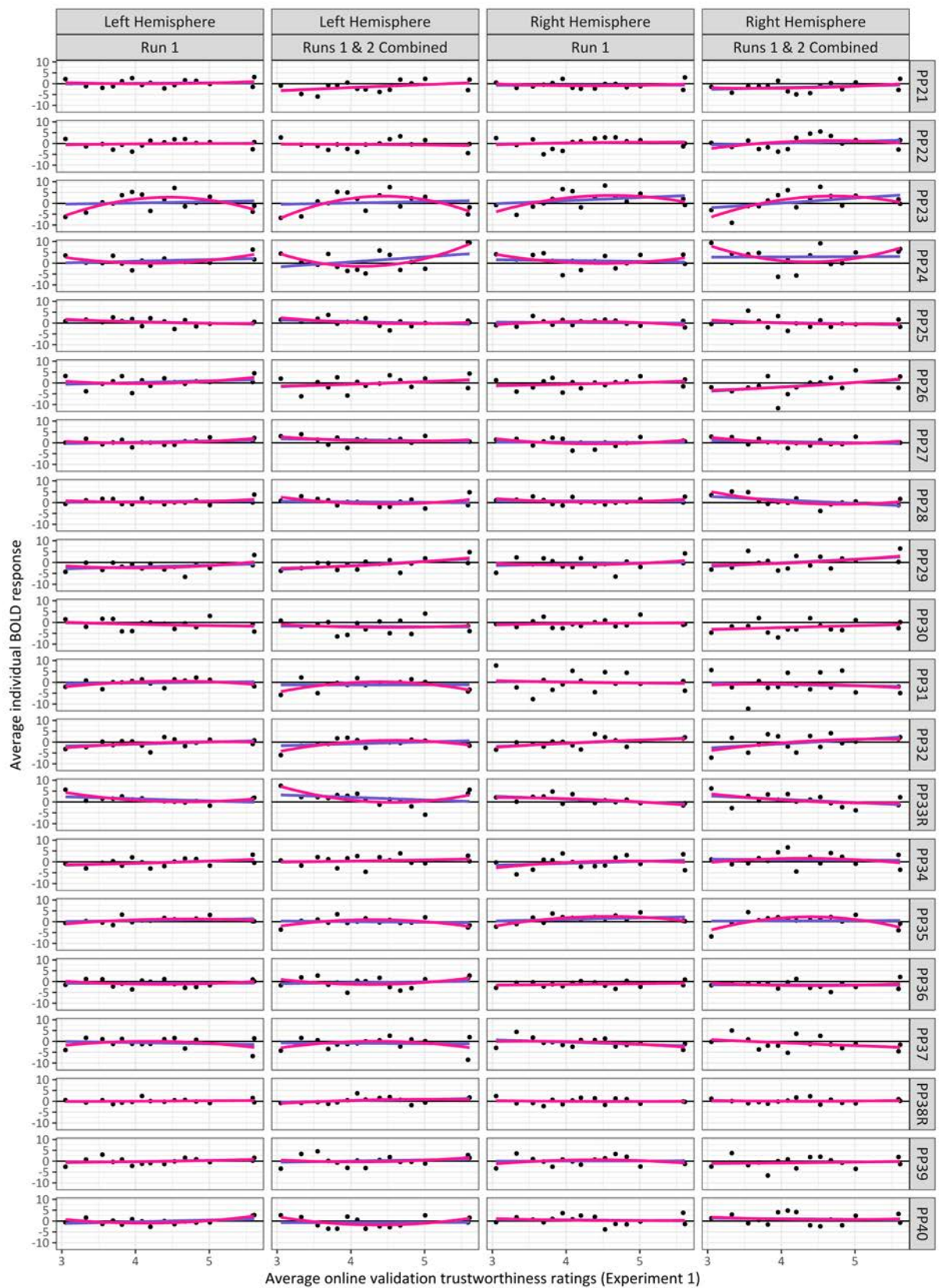


Figure 20: Male voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

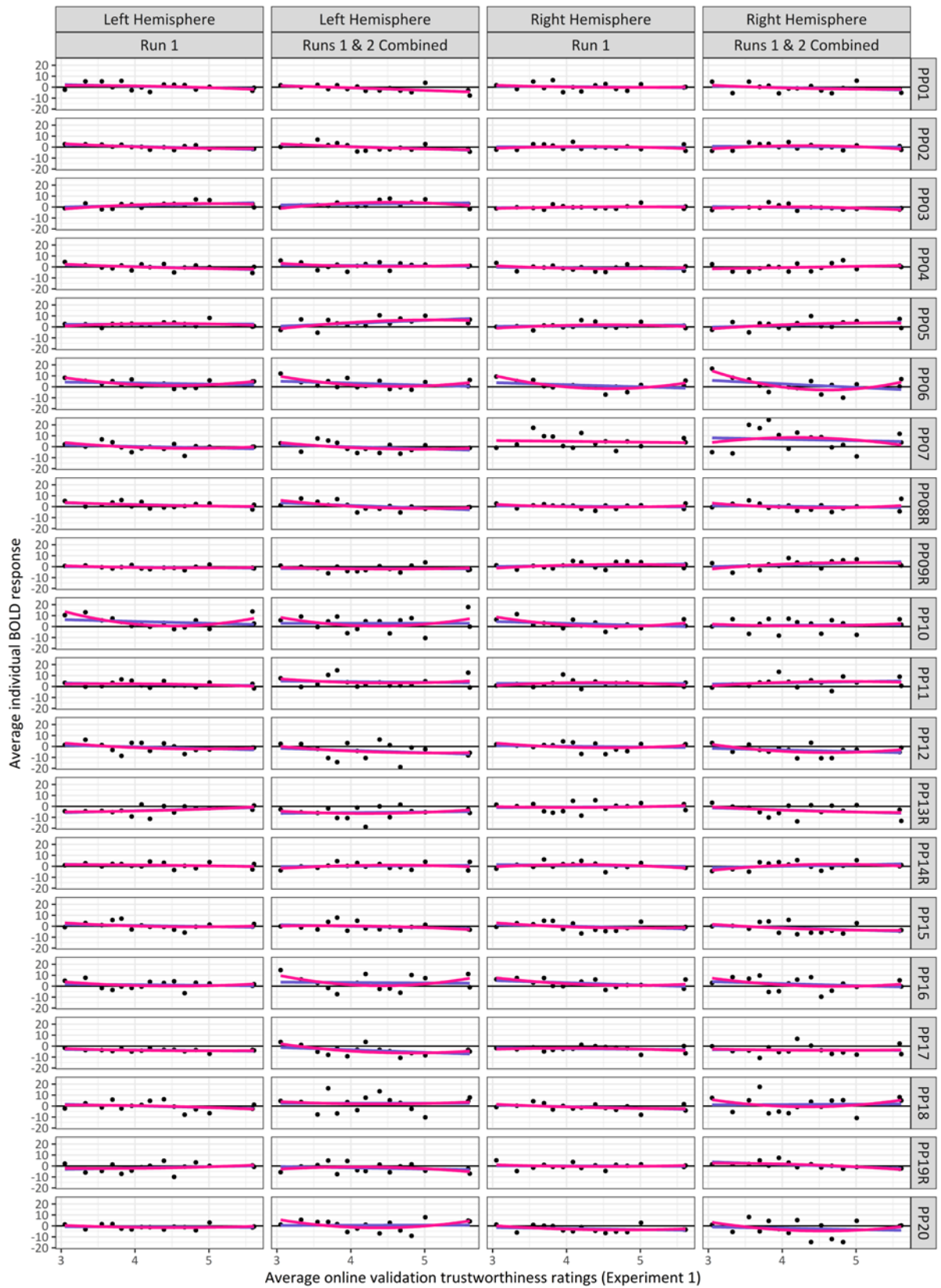


Figure 21: Male voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

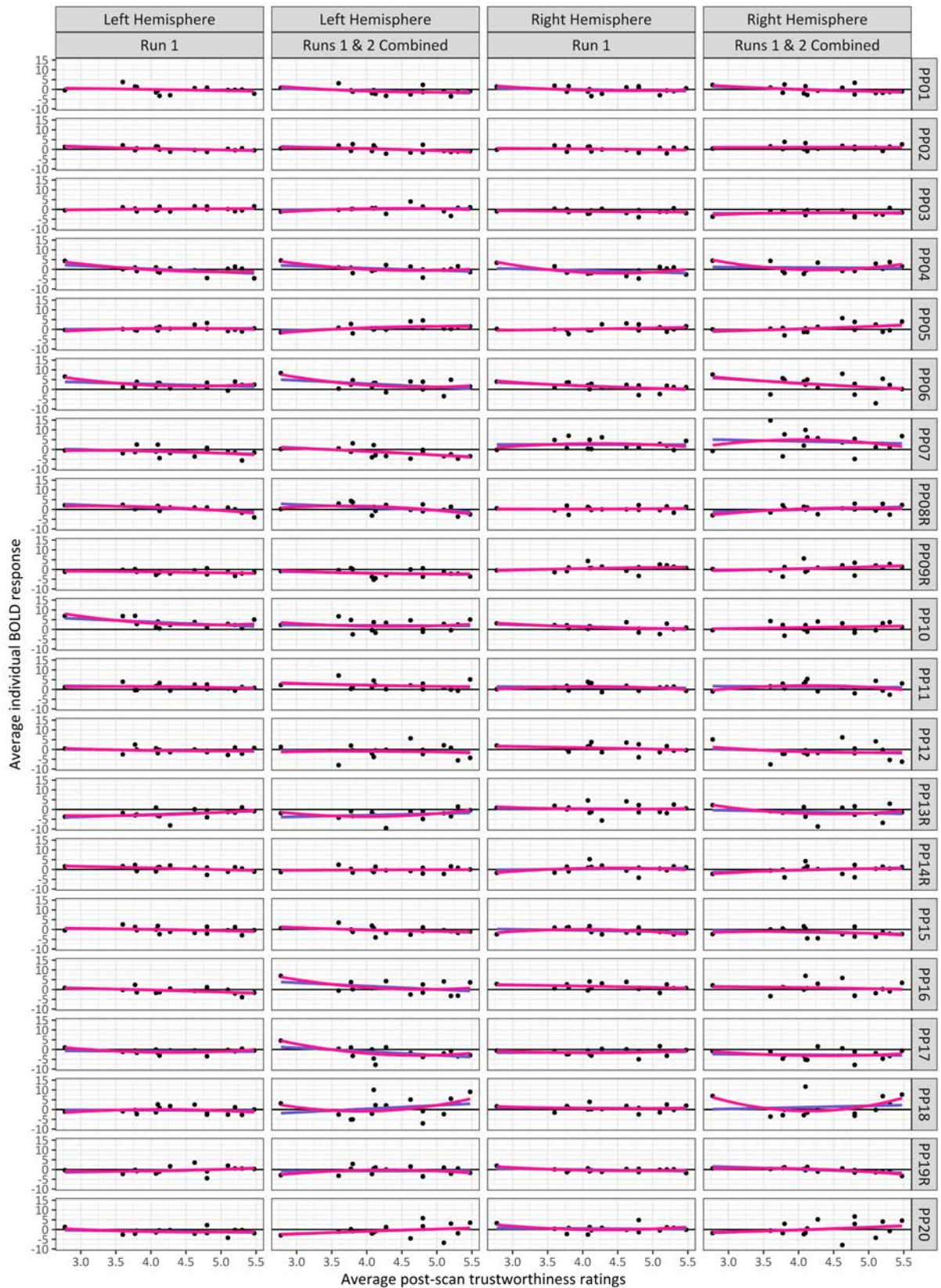


Figure 22: Male voices in the composite amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

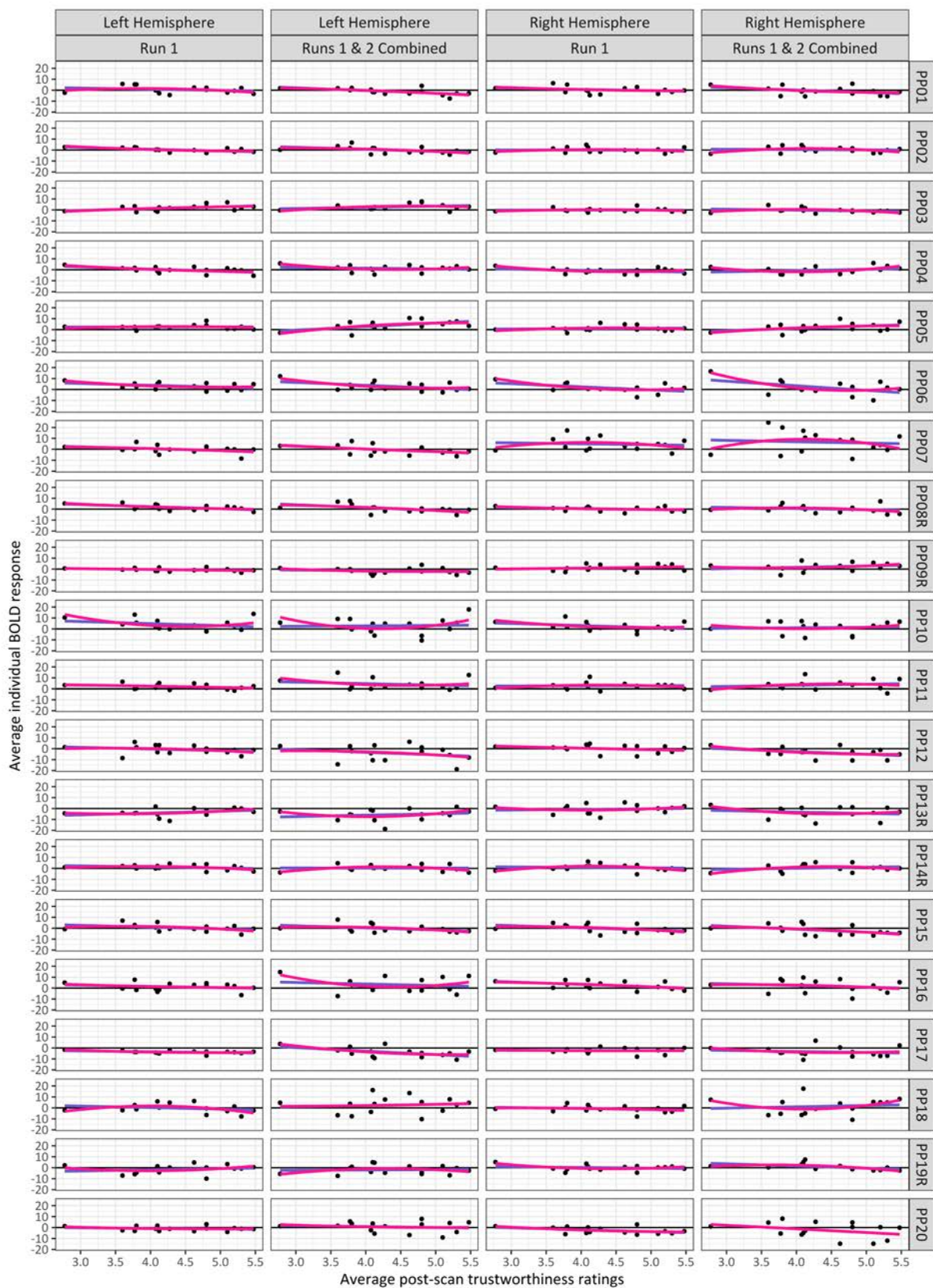


Figure 23: Male voices in the SF amygdala for individual participants in Experiment 2a – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

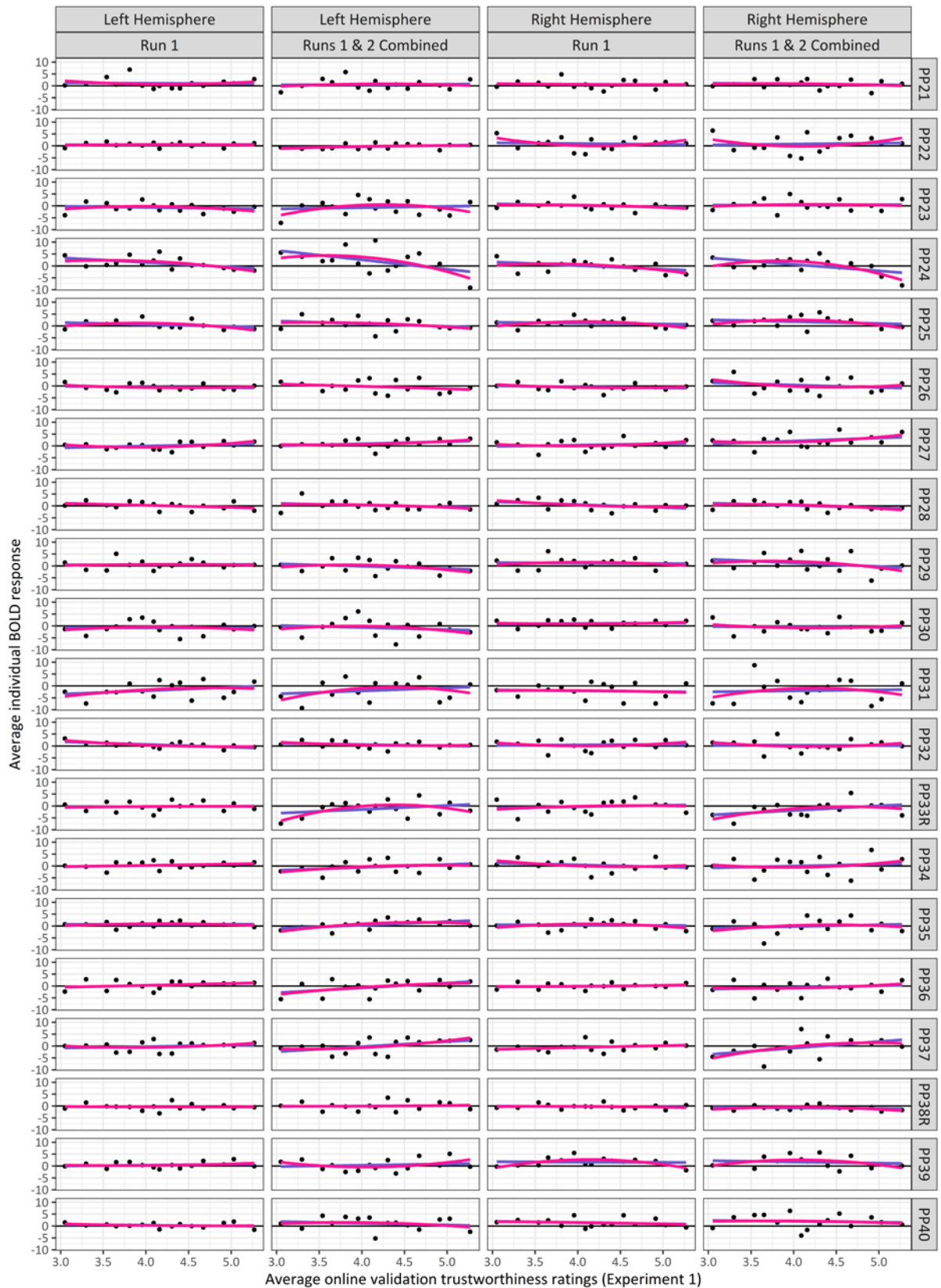


Figure 24: Female voices in the composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

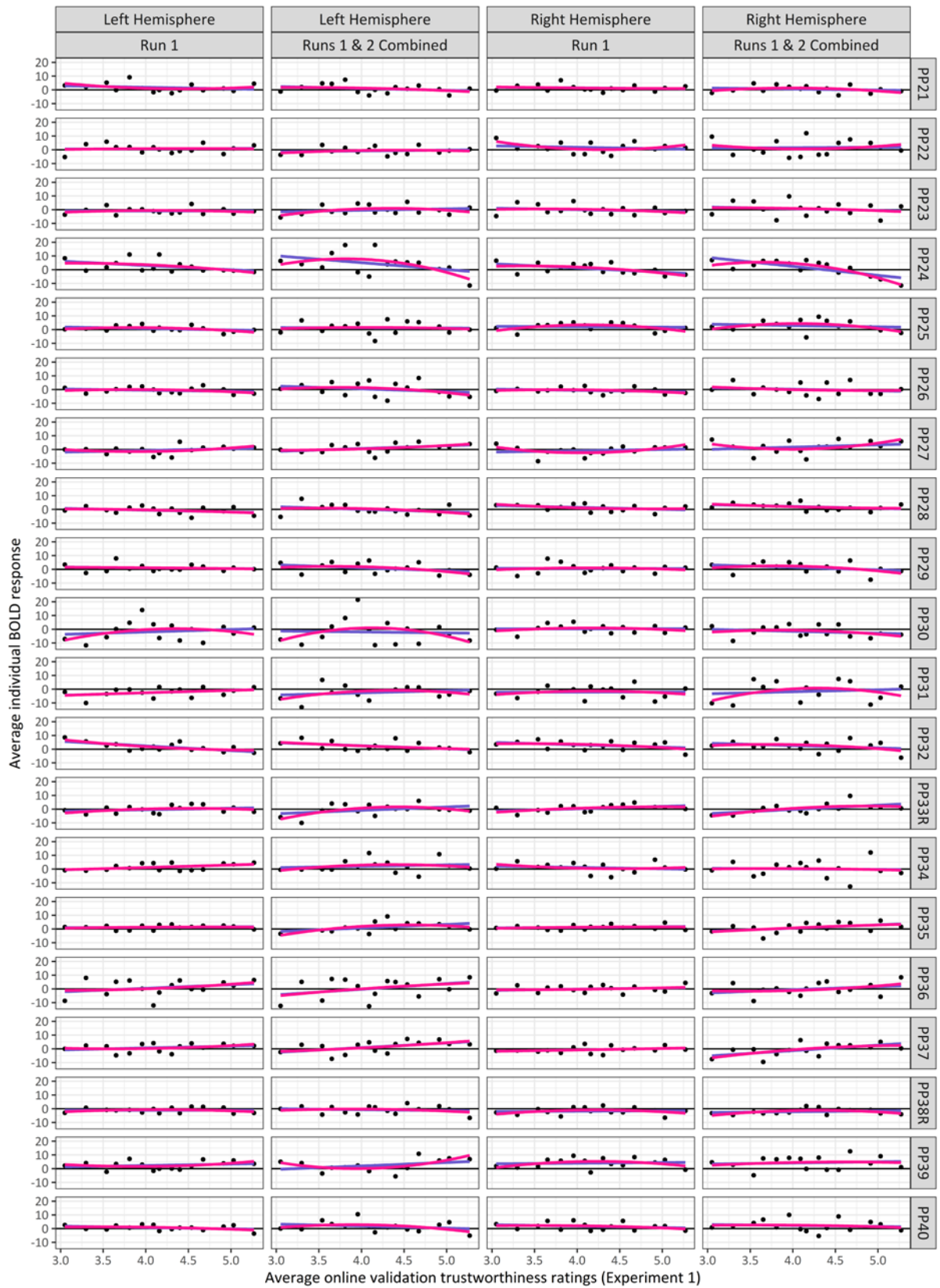


Figure 25: Female voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

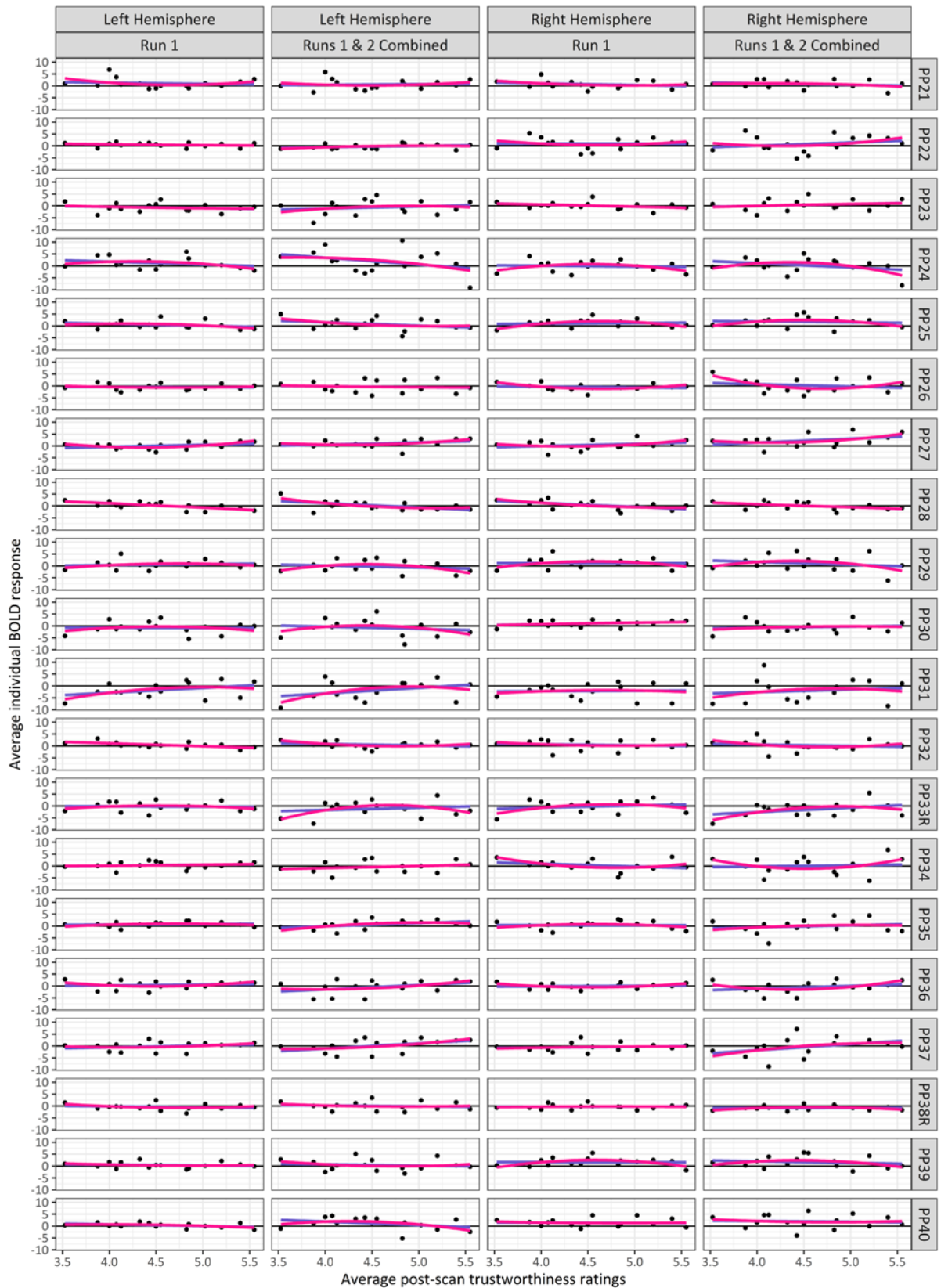


Figure 26: Female voices in the composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

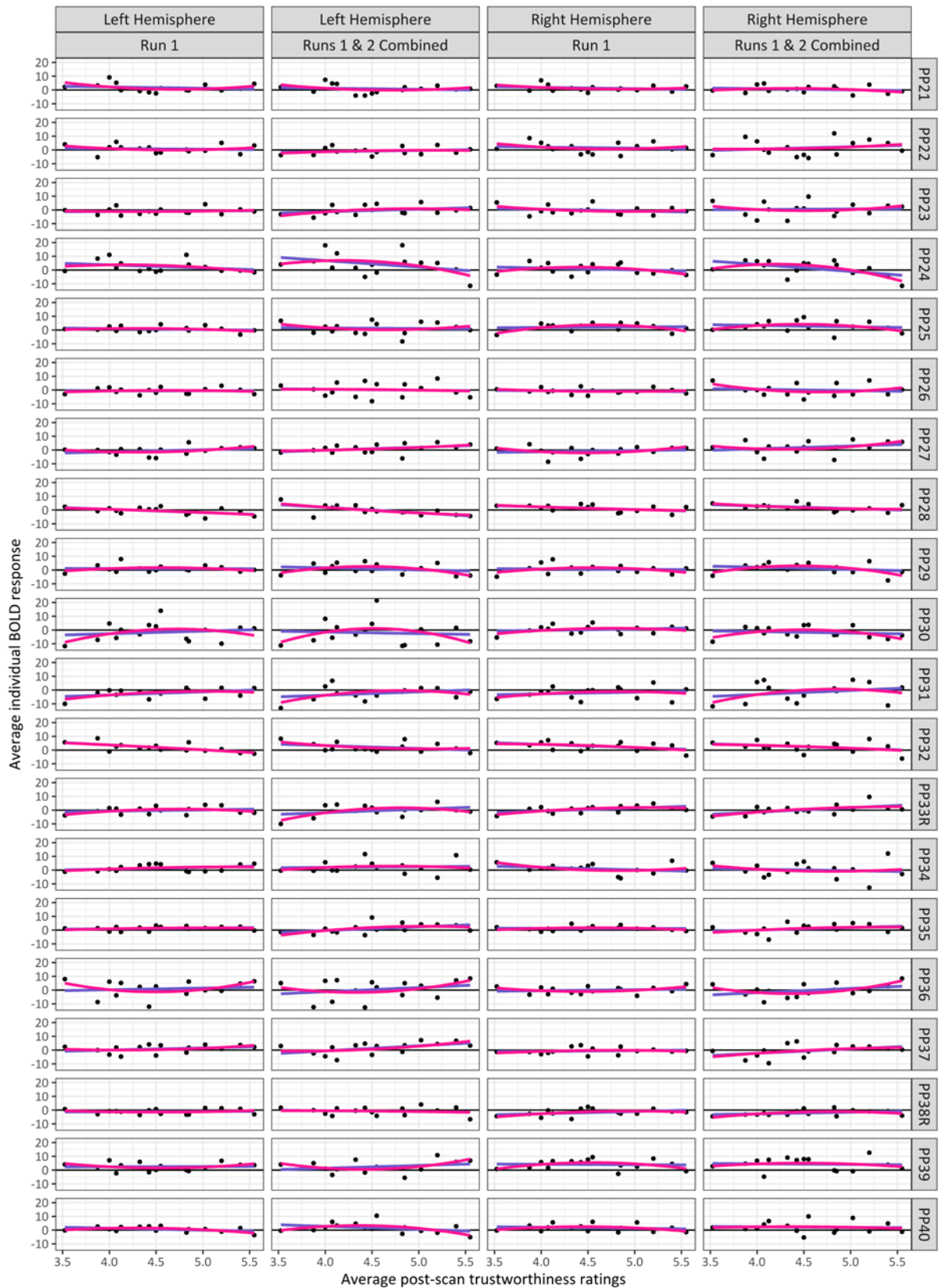


Figure 27: Female voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

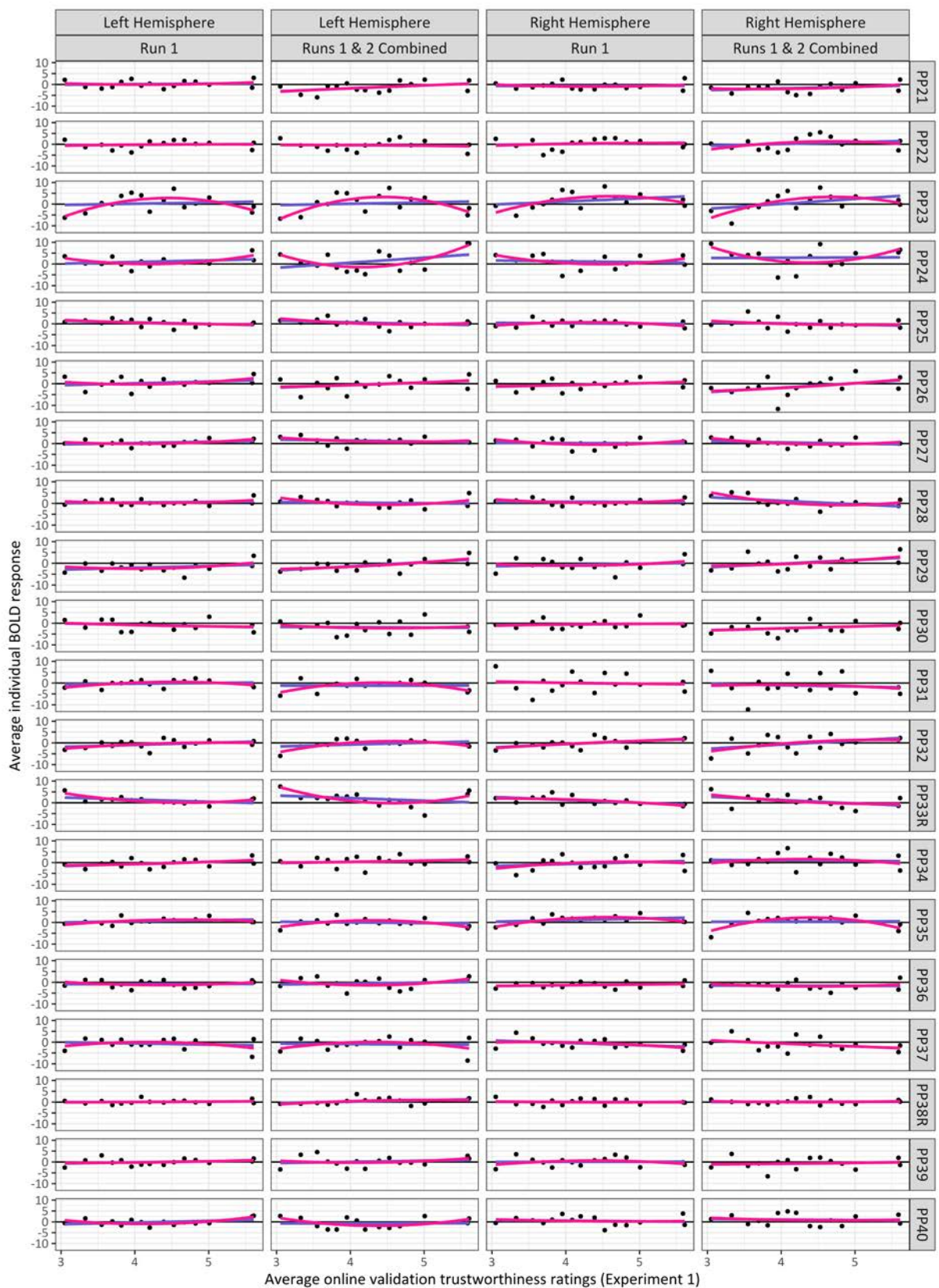


Figure 28: Male voices in the Composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

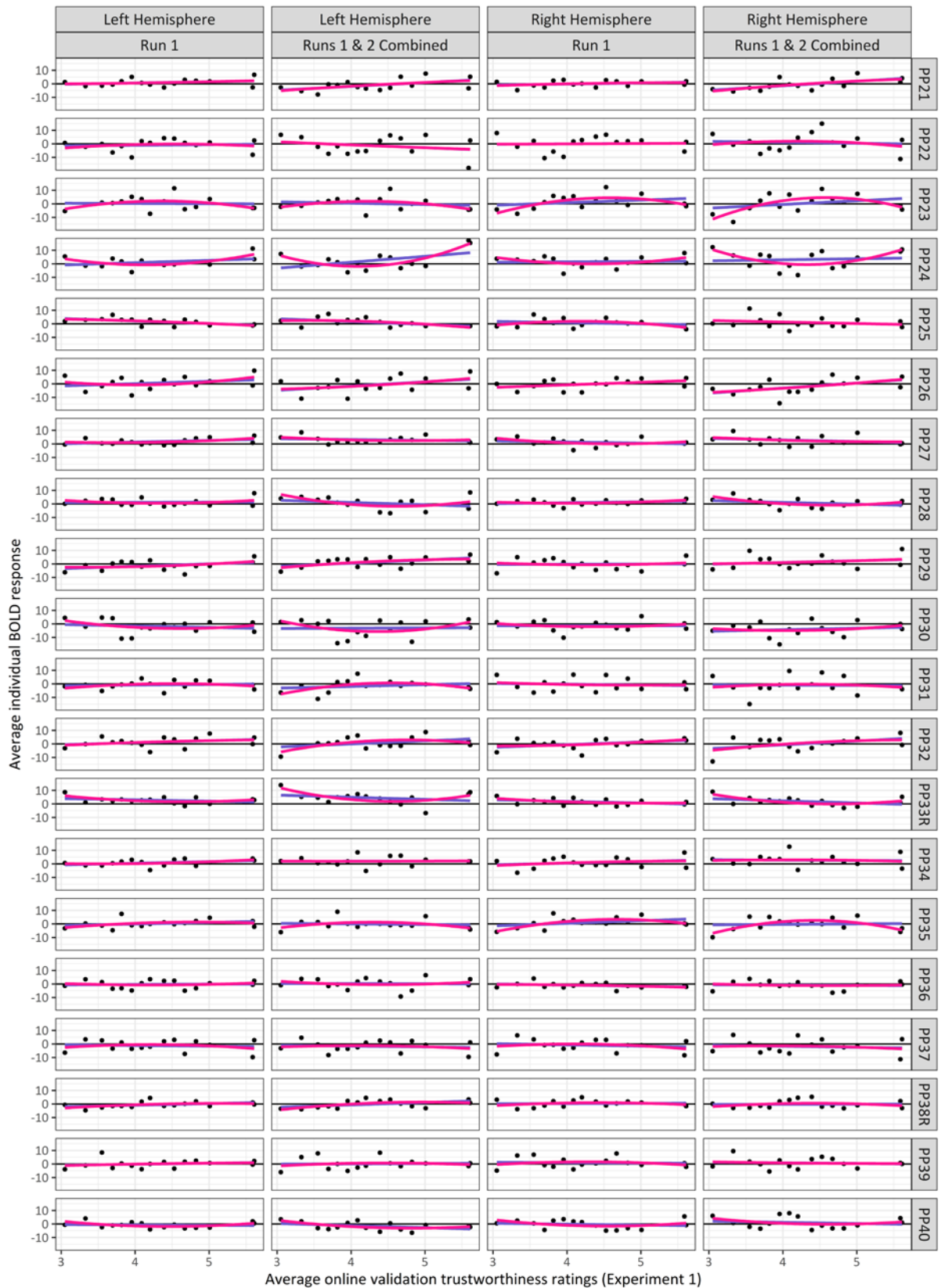


Figure 29: Male voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the online validation trustworthiness ratings (Experiment 1), separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

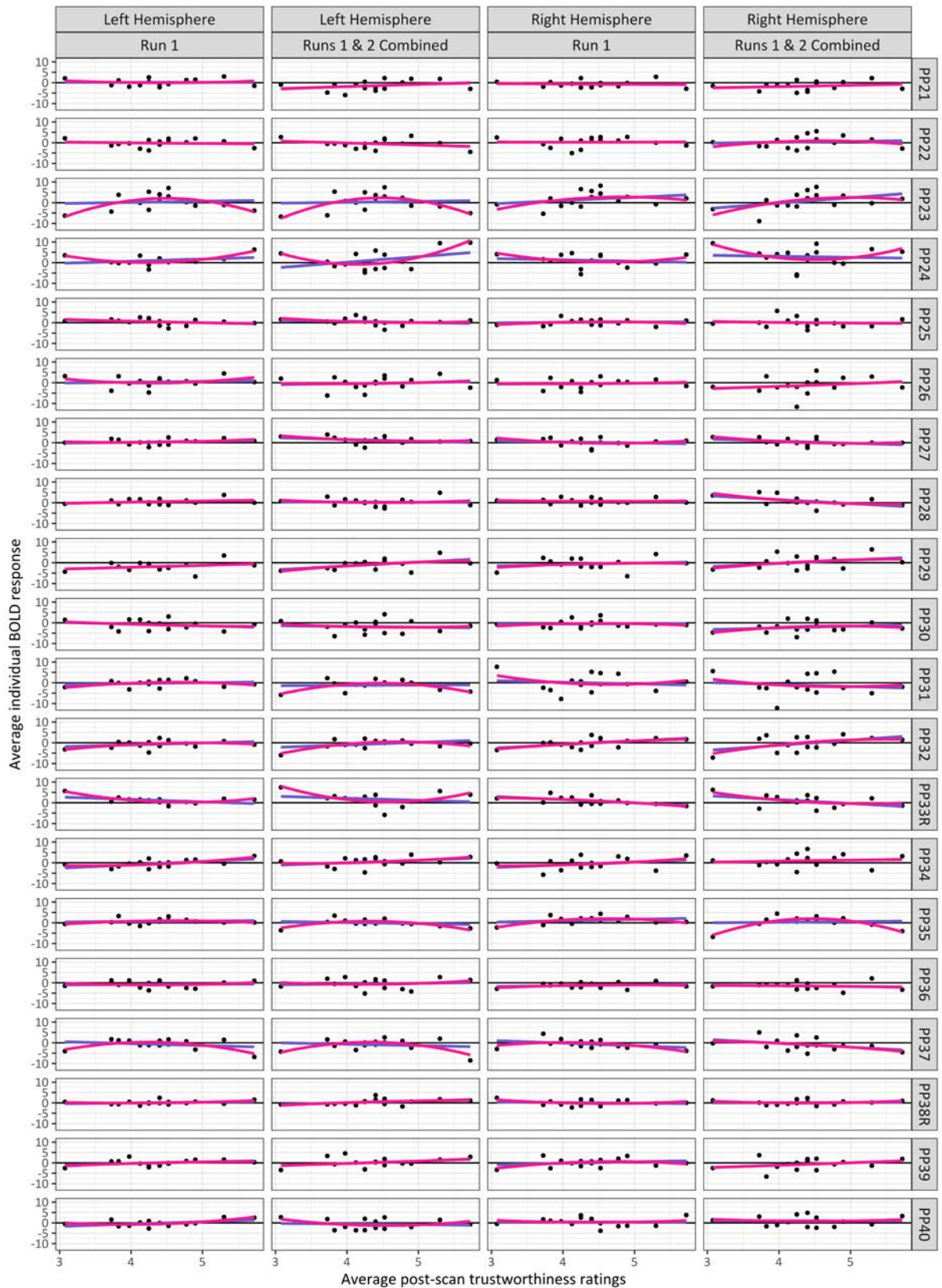


Figure 30: Male voices in the Composite amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.

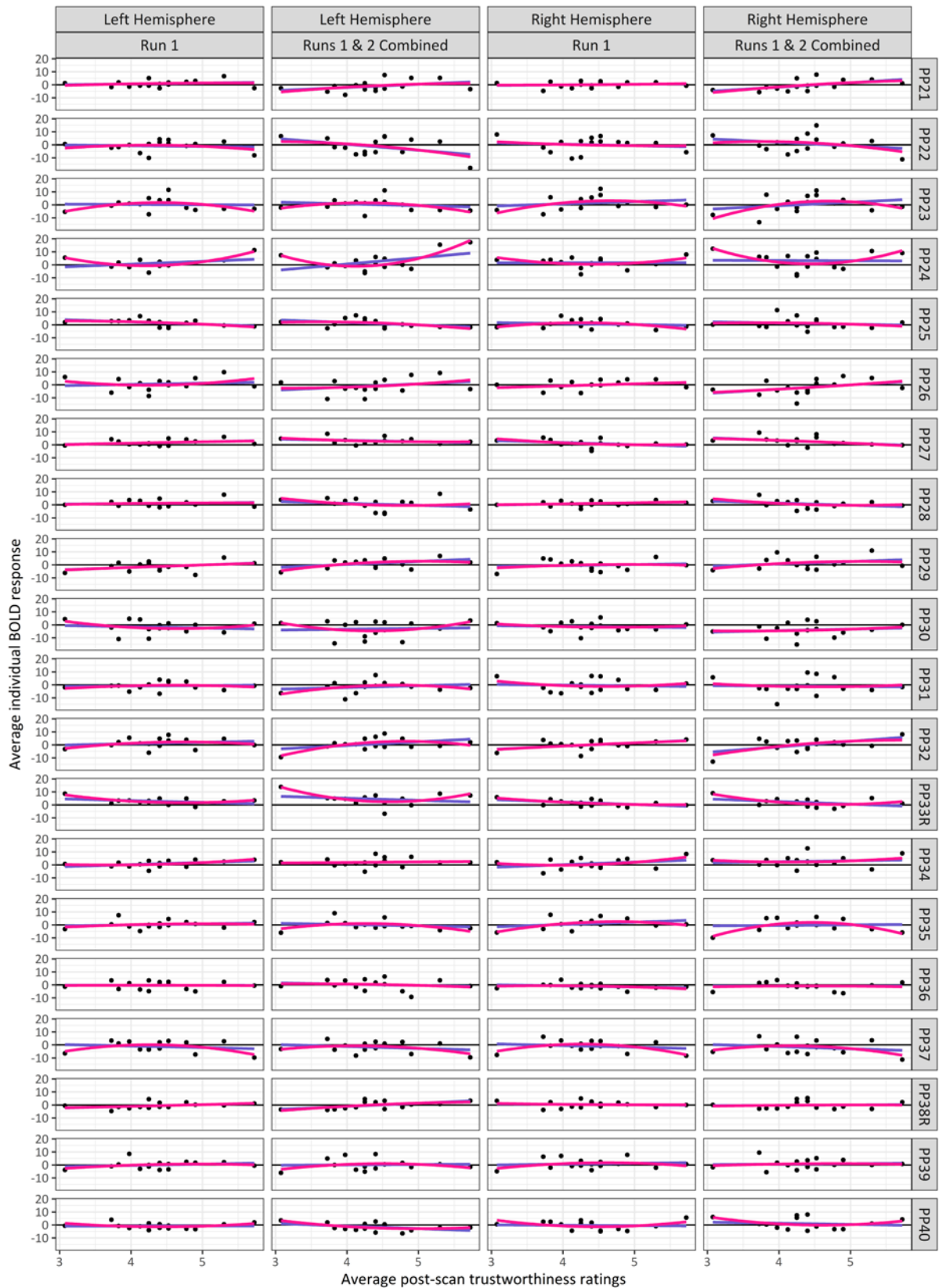


Figure 31: Male voices in the SF amygdala for individual participants in Experiment 2b – Average BOLD response in voxels in relation to the average post-scan trustworthiness ratings, separately for runs, and hemispheres. Linear (blue) and second-order polynomial (red) trendlines were added.